
# Chamberlain, James (2021) Hume's moral sentiments and Humean moral aliefs. PhD thesis, University of Nottingham.

# Hume's Moral Sentiments and Humean Moral Aliefs

**James Chamberlain**

Thesis submitted to the University of Nottingham for the

degree of Doctor of Philosophy

November 2020

## Abstract

*This thesis addresses the following two questions. How should we understand Hume's theory of the causes and nature of moral judgements? What can we learn from Hume in these regards?*

*I begin by reinterpreting Hume's theory of moral judgement. I argue that Hume claims, inter alia, that all moral judgements are sentiments, and that we experience a sentiment of approbation towards any token action or character trait of any type that we habitually associate with causing happiness to people around us. This interpretation easily reconciles Hume's claim that we consistently approve of justice because we sympathise with the happiness that it causes with his acknowledgement that many token acts of justice cause only unhappiness. Unlike most interpretations, mine does not entail that we ever correct moral judgements by adopting a 'common point of view'. I argue that these are positive features of my interpretation.*

*I then combine several features of Hume's theory, so understood, with recent theories of intuitive and associative mental states. I argue that typical moral judgements involve what Tamar Szabó Gendler calls 'occurrent aliefs': intuitive, associative mental occurrences with representational, affective, and behaviour-inducing content. Moral judgements are typically intuitively produced, as Jonathan Haidt argues. Contra Haidt, I argue that all paradigmatic wrongness judgements involve associations with harm, so that we need not posit fundamentally different kinds of wrongness judgement. I conclude by developing an 'emotive subjectivist' theory of the meanings and pragmatics of moral language. I argue that this coheres with several core features of Simon Blackburn's 'quasi-realist' metaethical expressivism.*

## Acknowledgements

First, I would like to thank my supervisors. Neil Sinclair has been a constant and invaluable source of help, encouragement, philosophical insight, metaethical debate, and general good-will over the last seven years. Every aspect of this thesis has been greatly improved by his input and guidance. Harold Noonan encouraged me to carefully connect together Hume's theses throughout his *Treatise*, which greatly aided me my interpretation of Hume's moral sentimentalism. Craig French helped me navigate the eddies and currents of various philosophical and psychological waters, and often provided clarity just where I most needed it. Many thanks to you all.

I would like to thank the Philosophy Department at the University of Nottingham, and all those who offered me support or guidance, or who discussed aspects of this thesis with me. This includes, but is not limited to, Andrew Fisher, Joseph Kisolo-Ssonko, Marcus Lee, Penelope Mackie, Jon Robson, Benedict Rumbold, Jonathan Tallant, and Christopher Woodard. My thanks also to Nottingham's postgraduates, for their many helpful comments.

I would also like to thank the organisers and participants of the 2019 'Recasting the *Treatise* III' conference, at the Institute of Philosophy, Hungarian Academy of Sciences, Budapest, for their invaluable discussion of a paper that became Chapter 6 of this thesis.

I am very grateful to both my parents, for all their support over many years of study, and to my father for his excellent proof-reading. Thanks to my daughter, Alice, for making the last four years of writing more fun and fulfilling than they could possibly have been otherwise. Most of all, I am eternally grateful to my wife, Katie, without whose constant love, support, and encouragement I could never have finished this thesis.

Part 1 of this thesis contains extracts from two papers:

Chamberlain, James (2017). Justice and the tendency towards good: The role of custom in Hume's theory of moral motivation. *Hume Studies* 43(1): 117-137.

Chamberlain, James (2019). Hume's emotivist theory of moral judgements. *European Journal of Philosophy*: 1-15.

# Contents

## References to Hume and Hobbes

Throughout this thesis, I quote those of Hume's texts in the bullet list below as at https://davidhume.org, edited by Amyas Merivale and Peter Millican. My references follow the conventions on that site: a code for the text, followed by (where relevant) book, section, or part and (in all cases) paragraph number, followed by a page reference to a widely used edition of that text, as listed in the bibliography.

The text codes are as follows:

- 'T' followed by 'SBN' refers to the *Treatise of Human Nature*
- 'E' followed by 'SBN' refers to the *Enquiry concerning Human Understanding*
- 'M' followed by 'SBN' refers to the *Enquiry concerning the Principles of Morals*
- 'P' followed by 'Bea' refers to the *Dissertation on the Passions*

References to Hume's *Essays Moral, Political, and Literary* are to the Miller edition, cited as 'EPML', followed by page number.

Hume's letters are cited as 'HL1', followed by a page reference to Greig (2011).

References to Hobbes's *Leviathan* are to the Curley edition, cited as 'L', followed by chapter number and paragraph number.

In all quotations, emphasis is original unless otherwise stated.

# Introduction

Towards the end of Book 3 of his *Treatise*, Hume summarises his metaethical 'system' as follows:

> When any quality, or character, has a tendency to the good of mankind, we are pleas'd with it, and approve of it; because it presents the lively idea of pleasure; which idea affects us by sympathy, and is itself a kind of pleasure (T 3.3.1.14, SBN 580).

This thesis represents an attempt to both understand Hume's system and to learn from it.

In Part 1, I attempt only to interpret and to understand Hume: I generally investigate the plausibility of his claims only for the purposes of charitable interpretation. I argue for a new interpretation of Hume's system, by which the summary above is to be understood as follows: A 'quality' or 'character' with 'a tendency to the good of mankind' is a token character trait, which is of a type that one generally associates with causing happiness to the possessors of such traits or to those people around them. Whenever one forms *any* idea of such a trait, then 'custom' or habit' ensures that one automatically forms an idea of the general type of pleasure that one associates with it (T 1.3.13.8, SBN 147). Any idea of such pleasure will be 'lively' to some extent: the idea will feel like it represents something *real*, even though the pleasure is not believed to exist. I call such ideas 'quasi-beliefs'. Any such quasi-belief in pleasure will produce at least some degree of genuine pleasure, via a process of 'delicate sympathy': a psychological mechanism by which we feel pleasure (or pain) at general ideas of pleasures (or pains) that are produced via customary association (T 3.3.1.8, SBN 577). And wherever one feels pleasure that has been caused in this way by a character

trait, then one experiences a consequent 'approbation' towards the token trait concerned (T 3.1.2.3, SBN 471). By my interpretation, Hume argues that *all* positive moral judgements are occurrences of approbation, produced in this way.

This argument allows Hume to endorse a thesis, which I call 'Generality':

GENERALITY: All ideas of typically beneficial or pleasing character traits habitually cause approbation, all ideas of typically harmful or displeasing character traits habitually cause disapprobation, and the strength of any moral sentiment is dependent only on the degree of happiness or unhappiness with which the type of trait is generally associated.

Hume never states Generality in these terms, but I will argue that he presents several arguments for this thesis, mainly in Book 3 of his *Treatise*. According to Generality, the moral sentiments are *uniform*:

UNIFORMITY: Any kind of evaluative or psychological response is *uniform* if and only if it is such as to respond in the same way towards all token character traits of any one type, regardless of how the responder is related to the person whose trait it is, or is affected by the particular effects of the token trait.

Approbation is a uniform sentiment, according to Hume, in that we experience a strong sentiment of approbation towards any token character trait of a generally pleasing

type, such as benevolence, whether or not we have any relationship with the benevolent person, and regardless of any particular effects of her benevolence.

Hume believes that there is very clearly a strong correlation between our moral approval of certain character traits and our expectations that such traits will cause happiness. His endorsement of Generality, and of the associated thesis of uniform sentiments, allows him to explain a range of what he takes to be otherwise perplexing counterexamples to this. Two of the most important examples can be summarised as follows:

> *Justice and other artificial virtues*: According to Hume, some types of character trait only cause happiness because we have, for non-moral reasons, developed various 'artificial' systems that benefit society. One such is 'justice', which includes the practice of repaying money that we have borrowed from others. Justice has a strong overall tendency to benefit society. However, there are some cases where repaying money causes *only* unhappiness to all concerned. Yet we still approve of the motive to be just in such cases.
>
> *'Virtue in rags'*: Hume sees a clear role for sympathy in the causes of our moral sentiments. Typically, we are pleased by acts of kindness, for example, because we sympathise with the beneficiaries of these acts. Hume thinks it obvious that sentiments of approbation have been caused, somehow, by sympathetic pleasures. Yet we approve of people who unsuccessfully try to be kind just as much as we do of those who succeed: 'Virtue in rags is still virtue' (T 3.3.1.19, SBN 584). In such cases, there are no beneficiaries with whom we may sympathise.

In both cases, I will argue that Hume explains why we approve of the seemingly anomalous token motives by appeal to Generality and to the uniformity of moral sentiments.

This is, so far as I know, a novel interpretation of Hume. Nevertheless, there are several widely recognised aspects of his theory of moral judgement that I endorse. It will be helpful to summarise some of the details of my interpretation here, all of which I will argue for in Part 1.

Hume's is a *sentimentalist* theory, by which moral judgements are to be explained primarily by reference to sentiments or passions, rather than by reference to reasoning or belief.

I argue that Hume understands *all* moral judgements as sentiments of approbation or disapprobation, rather than reasoned beliefs. Here, and throughout my thesis, I use the term 'moral judgement' to refer to what we would consider a mental state – what Hume calls a 'perception' – rather than to an utterance or verbalised moral evaluation (T 1.1.1.1, SBN 1). Sentiments are impressions and beliefs are ideas. I argue that Hume holds a 'Vivacity Thesis': that wherever we have a present impression of X, we cannot simultaneously form a believed idea (i.e. a 'vivid' idea) of X. I will argue that, given Hume's theories of impressions, ideas, beliefs, and the meanings of our utterances, this makes his theory an *emotivist* one, by which all moral utterances express moral sentiments.

Hume's is both a naturalistic and anti-realist theory. It is naturalistic in that it treats all its objects as scientifically explicable, if explicable at all: it has no place for non-natural objects or explanations. It is anti-realist in that it involves no facts or properties that could constitute the appropriate objects of putative moral beliefs. Hume's moral ontology consists of nothing more than ordinary, non-moral facts and properties in the world and our moral evaluations of some of these.

Hume's moral judgements have several important features, including the following:

- Approbation is a particular kind of pleasure. Disapprobation is a particular kind of pain. (T 3.1.2.3, SBN 471)
- They are directed primarily towards people's motives or character traits rather than towards their actions, although we may derivatively approve or disapprove of actions where we associate them with morally relevant motive types. (T 3.2.1.2, SBN 477)
- As pleasures and pains, they constitute our evaluations of character traits and motives, and so of actions or people, as 'virtuous' or 'vicious', 'good' or 'evil'. Evaluative sentiments cannot directly provide any motivational force, but they often indirectly provide motivational force by causing desires. (T 2.3.9.7, SBN 439)
- They are 'calm' sentiments, which means that their emotional feeling may be barely perceptible, although they may still be (indirectly) motivating. (T 2.1.1.3, SBN 276)
- The processes by which they are produced are very commonly mistaken for processes of reasoning, and they in turn are very commonly mistaken for evaluative beliefs. The desires which they cause are very commonly mistaken for beliefs about how we ought to act. (T 2.3.3.8, SBN 417)
- They are analogous to the sentiments in virtue of which we possess an aesthetic taste. They constitute our 'sense of beauty and deformity in action' (T 2.1.1.3, SBN 276)
- Approbation is a calm form of the passion of love. Disapprobation is a calm form of hate. (T 3.3.5.1, SBN 614)
- Approbation is caused by character traits of either generally useful or generally agreeable types. This is not a substantive distinction for Hume: useful or agreeable traits are morally important insofar as they are traits that cause (non-moral) happiness, and so please us by sympathy. (T 3.3.1.30, SBN 591)

A further important aspect of Hume's theory is his previously mentioned division of virtuous types of character traits – 'virtues' – into the 'artificial' and the 'natural' (T 3.2.1.1, SBN 477). Natural virtues are, roughly, motives to perform actions of types that we instinctively perform, such as the motive to care for one's child, where the actions so caused typically cause happiness to the actor or to those around her. Artificial virtues also typically cause happiness in this way, but they are motives to perform actions of types that we have no instinct to perform. We have learned to perform these actions for non-moral reasons, because they help us and our loved ones in some way. 'Justice' is Hume's paradigm of an artificial virtue, and (in the *Treatise*, at least) 'benevolence' is his paradigm of a natural virtue. Here, and throughout, I will adopt what I think is Hume's own approach, of using the names of virtue types purely descriptively, rather than as so-called 'thick' terms (Williams, 1985). 'Benevolence' is a term that implies no evaluation on the part of the term's user, but merely describes a type of motive. In the *Treatise*, Hume understands benevolence as, roughly, the desire to help those around oneself.

Throughout Part 1 of my thesis, aside from chapters 3 and 6, I will mainly consider the *Treatise*, although I will look to later works to help understand Hume's aims within this work. Several scholars, including Merivale (2019), Millican (2017), and Taylor (2015), argue for significant changes between Hume's *Treatise* and his later reworkings of the same topics. I will argue for something similar, concerning Hume's 'common point of view' thesis, in Chapter 6. To best understand Hume's arguments supporting his original *Treatise* Book 3 treatment of morality, we must first understand the theoretical underpinnings that he develops in Books 1 and 2.

In Part 2 of my thesis, I seek to learn from Hume, by using some aspects of his theory to attempt to answer questions or resolve debates within contemporary metaethics. One such

question is that of whether there is any one property that is common to all and only moral judgements. I consider one influential theory of moral judgements – 'Moral Foundations Theory' – which appears to suggest otherwise. I argue that, given some suitable theory of moral learning, such as Hume's, we may allow for a common, unifying property of moral judgements. I then argue for the stronger claim that all paradigmatic moral judgements are of a unified kind. Following arguments from Haidt (2001; 2012) and others, including Hume, I conclude that we should understand all paradigmatic moral judgements as produced by intuitive, nonconscious processes, rather than by processes of reflective, conscious reasoning. Drawing on Hume's thesis of Generality, I argue that paradigmatic moral judgements are, or include, associative and intuitively produced mental occurrences of the kind that Gendler (2008a, 2008b) calls 'occurrent aliefs'.

I conclude my thesis with a novel theory of the semantics and pragmatics of moral language, which is influenced by emotivist and expressivist theories as well as by Hume's theory. I suggest an 'opacity thesis', by which most moral thinking is intuitive and nonconscious, and I argue that this allows for the Humean claim that, whenever we contemplate morally salient objects, we experience moral aliefs, even if we cannot recognise this by introspection. I give reasons to think that our moral utterances derive their meaning by referencing our moral aliefs.

I then argue that this simple subjectivist theory can be rendered plausible by means of a thesis of strong indexicality, such that we may refer to our current moral aliefs, no matter the tense of the moral sentence uttered, and by a theory of the pragmatics of moral language. According to this theory, we implicate our desires to coordinate our moral judgements with those around us whenever we utter moral sentences. I suggest that the kinds of desires being implicated can be understood much as expressivists understand moral judgements, and I conclude by arguing that my emotive subjectivism can plausibly draw heavily on many of

Blackburn's (1998) 'quasi-realist' arguments to explain and vindicate moral language. I do not attempt a thorough vindication along these lines, but I give reasons to think that such a vindication is achievable.

Here is a summary of what is to follow:

In Chapter 1, I discuss the background to Hume's theory of moral sentiments, with a focus on the sentiment of approbation. I consider his theories of impressions, ideas, causal reasoning, and sympathy. I argue that he thinks that all ideas are perceptions that represent by being copies of that which they represent, whereas no impressions are representative in this way. I then argue that, because of this, we should understand Hume to implicitly endorse an extrinsic, causal account of the intentionality of passions, as several scholars have argued.

In Chapter 2, I consider Hume's distinctions between direct and indirect passions, and between calm and violent passions. I first argue that Hume sees a generally unappreciated but vital role for the formation of complex ideas in the production of four indirect passions: pride, humility, love, and hatred. This allows him to develop an account of their intentionality that is more coherent than many think possible. I then argue that calm passions are calm because they are caused by more generalised ideas than those that cause violent passions.

In Chapter 3, I consider Hume's theories of delicate sympathy and of motivation, and I argue for a novel interpretation of his theory of the causes of the moral sentiments, based on my arguments from chapters 1 and 2. I argue that approbation is a calm form of love, that disapprobation is a calm form of hatred, and that both are produced via processes of pre-reflective associations of ideas, and of our delicate sympathy with the quasi-beliefs so produced. I then look to Hume's later *Enquiry concerning the Principles of Morals*, and I argue that his theory of the causes of moral sentiments is, in all fundamental details, the same there as in the *Treatise*.

In Chapter 4, I consider Hume's discussion of justice. I survey a range of interpretations of Hume's treatment of justice, and I argue that they are all subject to insurmountable objections. I argue that Hume structured his *Treatise* so as to use the case of justice as evidence for Generality and his thesis of delicate sympathy.

In Chapter 5, I consider and reject some recent arguments that Hume allows for some moral judgements to be beliefs. I argue for Hume's Vivacity Thesis: that, wherever we have a present impression, we cannot hold a vivid idea in mind which differs from that impression only in its level of vivacity. This requires Hume to endorse an emotivist theory, by which moral utterances are expressions of either approbation or disapprobation.

In Chapter 6, I conclude my interpretation of Hume by reassessing his notion of a 'common point of view', which he claims we adopt whenever we moralise. I focus on the problem that this discussion is intended to resolve: that of satisfactorily explaining why our verbal assessments of character are less variable than most of the passions which result from contemplating people's characters. I argue that Hume's *Treatise* response to this problem is heavily influenced by a Hobbesian theory of value, which leads him to develop a complex account of our reasons to express only our uniform, moral sentiments when we publicly evaluate characters. In the moral *Enquiry*, to take up the common point of view is more simply to express our uniform sentiments via the use of moral language.

Having concluded my interpretation of Hume, I turn to contemporary metaethics. I ask what we can learn from some aspects of Hume's theories; notably, his thesis of Generality and his account of the associative causes of moral judgements.

In Chapter 7, I adopt one of the most influential accounts of moral judgements: Moral Foundations Theory (MFT). This entails that moral judgements are produced by intuitive, associative processes, but it also suggests to some readers that moral judgements cannot be plausibly understood as a single kind of mental state or occurrence. I will argue that typical

moral judgements plausibly have a unifying feature, in virtue of the habitual way in which we learn to respond morally to certain action and character kinds. Therefore, we cannot infer, from anything entailed by MFT, that paradigmatic moral judgements may not form a unified psychological kind.

In Chapter 8, I argue for a 'moral alief theory', by which paradigmatic moral judgments are or include aliefs. I argue that this is compatible with a plausible explanation of the role of reflective thought and reasoning in the formation of many moral judgments.

In Chapter 9, I develop a theory I call 'emotive subjectivism': a relatively simple subjectivist theory of the meaning of moral terms, coupled with a pragmatic account of moral language.

In Chapter 10, I consider reasons to rethink some aspects of ordinary moral language, as typically understood by contemporary metaethicists. I consider what I take to be our best available approach to explaining and vindicating ordinary moral language: quasi-realist expressivism. I argue that emotive subjectivism can use similar strategies to explain and vindicate ordinary moral language, as we should understand it.

## 1. The Difference Betwixt Feeling and Thinking

Towards the end of his *Treatise*, Hume argues that our moral sentiments are pleasures and pains that 'can proceed from nothing but our sympathy with the interests of society' (T 3.3.1.12, SBN 580). Four sections later, he introduces the notion that approbation 'is nothing but a fainter and more imperceptible love' (T 3.3.5.1, SBN 614).[1] It now appears that approbation is a 'calm' form of the passion of love, so that it differs from non-moral love only in its absence of any 'violent' emotional feeling (T 2.1.1.3, SBN 276). Yet in Book 2, 'Of the Passions', Hume discusses the causes of the 'indirect' passion of love in some detail, and he appears to treat approbation as a distinct passion from love (T 2.1.1.4, SBN 276). In Book 2, Hume classifies approbation among our sentiments of taste: it is that sentiment in virtue of which we possess a 'sense of beauty…. in action' (T 2.1.1.3, SBN 276).

Despite claiming that we approve of actions at T 2.1.1.3, Hume says in Book 3 that wherever we approve of actions, this is only because we consider them as 'signs of… motives' (T 3.2.1.4, SBN 478). In such cases, the 'ultimate object' of approbation is always the 'motive' behind the action, such as a desire to help others, or to repay a loan (T 3.2.1.2, SBN 477). We may approve of any 'quality of the mind' that makes a 'character' 'naturally fitted' to cause happiness, by being useful to others or to oneself, or by being agreeable to others or oneself (T 3.3.1.30, SBN 591). All such qualities are character traits. These traits are typically motives, but they may be other kinds of traits too, such as wit (T 3.3.4.8, SBN 611). We approve of all and only those traits with a 'tendency to the good of mankind' (T 3.3.1.10, SBN 578).

[1] As Hume does, I will often concentrate on approbation, and assume that all relevant theses and arguments also apply, *mutatis mutandis*, to disapprobation.

If approbation is a sentiment of taste, then it must be a calm passion, because all such sentiments are calm, according to Hume (T 2.1.1.3, SBN 276).[2] If it is a form of love, then it must be an indirect passion, as love is (T 2.1.1.4, SBN 276-7). Hume calls passions 'indirect' where they are caused by relatively complex psychological processes, unlike 'direct' passions, such as desires, which Hume thinks have relatively simple causes (T 2.1.1.4, SBN 276).[3] Love, hatred, pride, and humility are each caused via a notoriously complex process, involving a 'double relation of ideas and impressions' (T 2.1.5.5, SBN 286).

Over this and the next two chapters, I will examine three core theses within Hume's theory of approbation, which may be summarised as follows: (1) Approbation is a calm form of the 'indirect' passion of love; (2) Approbation is a sentiment of taste, in virtue of which we come to value useful and agreeable kinds of character traits; (3) Approbation is in all cases caused, via sympathy, by an idea of a character trait with a 'tendency to the good of mankind'. I will argue that these claims can almost entirely be reconciled, so that Hume has a significantly more coherent account of approbation than is typically believed. In Chapter 3, we will see that Hume understands approbation as a faint or 'calm' form of love, *and* as a pleasing sentiment of taste, caused via sympathy with an idea of happiness.

These are, I believe, important points. Nevertheless, mine is not an entirely new position. Somewhat like Árdal (1966, 116) I argue for 'a close analogy' between approbation

[2] I will generally use the term 'sentiment' to mean a calm passion, although Hume does not consistently use the term in this way. 'Sentiment' is not a technical term, and it may refer to any kind of preference, opinion, or view. Hume sometimes calls a belief a 'sentiment' (e.g. T 1.2.2.3, SBN 30). He sometimes calls a violent passion a 'sentiment' (e.g. T 1.3.10.10, SBN 631; T 2.2.5.2, SBN 358). He often uses the term as a synonym for 'taste' (e.g. T 1.3.8.12, SBN 103).

[3] Hume distinguishes desires from aversions, but I will often refer to both as 'desires'. What Hume calls an 'aversion' may be understood as a desire to avoid something.

and love. Somewhat like Baier (1991, 135), I argue that 'moral evaluations are general', unlike the more 'particular' evaluations of love. I believe that my interpretation reconciles theses (1), (2), and (3), as previous interpretations cannot. However, I cannot hope to argue against each interpretation individually. A very brief survey of available opinion on the relation between love and approbation alone will demonstrate the range of different interpretations.

Árdal (1966; 1977) argues that approbation is a kind of indirect passion, closely analogous to love, which is made calm by the nature of its causes. Several others, including Brown (2001) and Mercer (1972), follow Árdal in this regard. Cohon (2008, 179) argues that approbation is an indirect passion, somewhat like love, but not as similar as Árdal suggests. Korsgaard (1999, 9) claims that 'Hume thinks that virtue and vice are intimately related to love and hate, but he is a little unsettled about what exactly the relationship is'.

Several other scholars deny that Hume argues for any significant connections between love and approbation. Baier (1991, 134) thinks that Árdal's view overstates the similarities between love and approbation, and that approbation is not an indirect passion. Kemp Smith (1966, 167) claims that approbation is a direct passion. Schauber (1999) argues that love is insufficiently motivating to have any close relation to approbation. Moreover, she claims that the causes of love are unrelated to the sympathetic causes of approbation. Garrett (2002, 193) claims only that Hume argues for 'distinctively moral impressions'. Carlson (2014) and Loeb (1977) argue that moral sentiments are *sui generis* calm passions that are neither direct nor indirect passions.

On my reading, Hume sees approbation, as felt towards others, literally as a calm form of love. Admittedly, it is not always possible to reconcile everything that he says. For Hume, love and hatred are always other-directed passions, and there is not strictly any such thing as self-love or self-hatred (T 2.2.1.2, SBN 329-30). Yet, for reasons that are not entirely

clear, approbation and disapprobation *may* be self-directed. This complication aside, I will argue that Hume consistently understands approbation as a calm form of love. I will conclude this argument in Chapter 3 but, first, I must address the background to Hume's theory of approbation. That will be the main topic of this chapter. Then, in Chapter 2, I will consider those elements of his theory of the passions that are most relevant to his theory of approbation.

In §1.1, I examine Hume's several brief summaries of the causes of approbation, which at least suggest that any token character trait of any generally useful or agreeable type will cause approbation. This is an interpretative thesis that I will be arguing for throughout my discussion of Hume. §1.2 addresses the core elements of Hume's theory of impressions and ideas, and those of his theory of causal reasoning, both of which are integral to his theory of sympathy. In §1.3, I discuss Hume's theory of sympathy, and I introduce his taxonomy of the passions. In §1.4, I consider the way that ideas represent their objects, according to Hume, and I address his implicit theory of the intentionality of passions.

### 1.1. Hume on the causes of approbation

Hume provides several summaries of the causes of the moral passions, including the following: 'virtue is distinguished by the pleasure, and vice by the pain, that any action, sentiment or character gives us by the mere view and contemplation' (T 3.1.2.11, SBN 475). By this, I propose that he means that any relevant action, sentiment or character will cause us to experience a moral passion before we have a chance to reflect on its likely effects. I propose this because several of Hume's other summaries of the causes of the moral sentiments suggest the same thing.

Consider Hume's initial, brief summary of (what we would now call) his metaethical position, in Book 2. There, he claims that 'certain characters and passions, *by the very view*

*and contemplation*, produce a pain, and others in like manner excite a pleasure… To approve of a character is to feel an original delight *upon its appearance*.' (T 2.1.7.5, SBN 296, my emphasis). This passage again suggests that moral sentiments occur *whenever* we encounter the relevant character traits, as soon as they appear to us, and regardless of any of our particular beliefs about them or their effects. Similarly; '[e]very quality of the mind is denominated virtuous, which gives pleasure *by the mere survey*' (T 3.3.1.30, SBN 591, my emphasis). Consider too Hume's summary of his aims for the second and third parts of Book 3: to answer the 'simple question, *Why any action or sentiment upon the general view or survey, gives a certain satisfaction or uneasiness*' (T 3.1.2.11, SBN 475).

In T 3.3.1.30 (SBN 591), Hume argues that those character traits which cause approbation are those that are 'naturally fitted' to be useful or agreeable, either to the person whose trait it is or to those around her. Typically, Hume talks of token objects being 'fitted' to cause an effect when they are tokens of types which are generally such as to cause that kind of effect. For example, a set of fortifications are 'fitted to attain their ends' where they are of kinds that we believe will successfully repel invaders, if required (T 2.3.10.5, SBN 450). In his discussion of approbation, 'naturally fitted' cannot mean 'non-artificially fitted', because, as we will see in Chapter 4, Hume believes that we approve of many character traits that *are* artificially developed to be useful. I take it that 'naturally' is intended to be 'opposed to rare and unusual', which Hume thinks is the most 'common' meaning of the word (T 3.1.2.8, SBN 474). Presumably, Hume means that we approve of traits of types that typically cause happiness.

This all at least suggests a theory that, as soon as one understands any character trait to be a token of any type that generally causes happiness, one will immediately experience approbation, before one has time to reflect on its particular consequences. Even if a token motive of benevolence, for example, ultimately produces no happiness, we will be very likely

to ‘conceive’ it ‘under the general notion’ of benevolence: we will be very likely to identify it *as* a motive of benevolence (T 2.3.6.2, SBN 424). Hume seems to be suggesting that we will approve of any motive of benevolence, as soon as we categorise it as such, merely because benevolence is ‘naturally fitted’ to cause happiness. I will argue that approbation is always caused by ideas of traits that are taken to be naturally fitted to cause happiness.

Admittedly, at T 3.3.1.30 (SBN 591), Hume appears to claim that we may approve of some token characters just because they are immediately agreeable to us, so that we might feel approbation towards someone because her witty comment makes us laugh, for example (see also T 3.3.1.27, SBN 589-90). I will return to this in Chapter 6, where I will argue that Hume ultimately denies that moral sentiments can be caused in this way. Indeed, at T 3.3.1.30, he has just stressed that we need not be *personally* pleased to approve of traits like wit: ‘We approve of a person, who is possess'd of qualities immediately agreeable to those, with whom he has any commerce; tho' perhaps we ourselves never reap'd any pleasure from them’ (T 3.3.1.29, SBN 590; see also M 8.15, SBN 267). The ‘principle of *sympathy*’ ensures, somehow, that we approve of all token traits that we take to be of useful or agreeable types, just because these types of traits generally cause pleasure (T 3.3.1.29, SBN 590).

In reading Hume in this way, I agree with those, such as Darwall (1994, 71) and Reed (2016), who argue that he distinguishes agreeable from useful traits only by the different ways in which they cause non-moral pleasures, so that he takes both useful and agreeable traits to produce approbation via our sympathy with these non-moral pleasures. Any trait that is naturally fitted to be useful or agreeable may cause approbation, and any trait that causes approbation does so because it is naturally fitted to be useful or agreeable.

With this in mind, consider Hume’s brief argument that the ‘good qualities of an enemy are hurtful to us; but may still command our esteem and respect’, because it is ‘only when a character is considered *in general*, without reference to our particular interest, that it

causes such a feeling or sentiment, as denominates it morally good or evil' (T 3.1.2.4, SBN 472, my emphasis). If one enemy pleases another by her generosity, we may be violently pained by this, but we may nevertheless approve of her simply because she is generous. This suggests that wherever we identify someone's motive as a generous one, then it will produce a calm sentiment of approbation, simply because generosity is generally pleasing to us.

Throughout my discussion of Hume, I will argue that, just as the above passages suggest, he believes that we approve of all traits that we habitually associate with causing happiness, regardless of our beliefs about their particular effects. There is some intuitive appeal to this view. We generally do love people for all their many particular quirks and attributes, and we generally do approve, or try to approve, of all similarly generous people to the same extent, just because they are generous. Before we can be certain that it is Hume's view, however, we must understand his theories of ideas, reasoning, passions, and sympathy. We must understand how Hume thinks that sympathy can cause us to feel approbation towards any trait of a generally pleasing kind, as soon as we identify it as a trait of that kind.

In §1.2, I begin my examination of this topic, by considering Hume's theories of impressions, ideas, and causal reasoning.

### 1.2. Impressions, ideas, and causal reasoning

Hume calls all mental objects 'perceptions', and all perceptions other than ideas 'impressions' (T 1.1.1.1, SBN 1). Impressions include sensory perceptions, feelings, passions, sentiments, pains, and pleasures. Ideas are, roughly, our thoughts, beliefs, and memories. Hume claims that we will all recognise the difference between impressions and ideas, which is simply 'the difference betwixt feeling and thinking' (T 1.1.1.1, SBN 2). To 'feel', according to Hume, is to directly experience something, rather than to think about or remember that thing. He can therefore claim – as indeed I think he does – that we feel

approbation, without meaning to suggest that approbation is experienced in anything like the way that a passionate feeling, like those of joy or wonder, are felt.

As Beebee (2006, 15) observes, Hume treats the mind as 'a kind of natural, quasi-Newtonian system', in which a relatively small set of principles underpin all the many associations and interactions between perceptions. As a careful observer of this system, Hume tries to set out in clear detail what these principles are, although he doubts that we could ever hope to find any explanation for the principles themselves.

The first of Hume's principles of the mind is that '*all our simple ideas in their first appearance are deriv'd from simple impressions, which are correspondent to them, and which they exactly represent*' (T 1.1.1.7, SBN 4).[4] This claim is now commonly called Hume's 'Copy Principle' (e.g. Garrett 2002, 41). As Garrett (2002, 49) notes, Hume's evidence for this principle is firmly empirical: experience convinces him that it is consistently the case (T 1.1.1.8, SBN 4-5). Simple ideas are ones that 'admit of no distinction', so that they cannot be analysed, such as an idea of a certain shade of red (T 1.1.1.2, SBN 2). We can form *complex* ideas of things that we have never experienced, but the Copy Principle requires that all the simple ideas that make up our complex ideas are copies of previous impressions. I have never seen a minotaur, but I have seen bull's heads and people's bodies before, and I combine my ideas of these into the complex idea of a minotaur.

Hume claims that ideas fundamentally differ from impressions only in their lower levels of 'vivacity' (T 1.1.1.3, SBN 2). He gives an example: 'That idea of red, which we

[4] Hume allows for a 'singular' exception to this 'general maxim', that is 'scarce worth our observing' (T 1.1.1.10, SBN 6). This is that someone who has seen many shades of one colour (blue, in Hume's example) would be able to form an idea of a sufficiently similar shade of which they have never had an impression. Hume seemingly allows that, in this very unusual case, the relation of resemblance between the ideas involved somehow leads to the formation of a new but very similar idea.

form in the dark, and that impression, which strikes our eyes in sun-shine, differ only in degree, not in nature' (T 1.1.1.5, SBN 3). He uses a 'variety of terms' to describe the feeling of vivacity, including 'liveliness', and 'force', although he admits, in a late addition to the *Treatise*, to using these in an 'unphilosophical' manner (T 1.3.7.7, SBN 629).[5] He stresses, in both his *Treatise* and in his later *Enquiry concerning Human Understanding*, that he is referring to a feeling properly called 'belief' (T 1.3.7.7, SBN 629; E 5.12, SBN 48-49). I shall henceforth assume that all the terms in question refer to the same property, which I shall generally call 'vivacity'. Although detailed interpretations of vivacity vary, I take it that any increase in a perception's vivacity involves (or, perhaps, is) an increase in the extent to which it seems really present (e.g. Boehm 2013; Dauer 1999; Waxman 2003). Impressions are maximally vivid, so that we take them to be immediately and presently real, rather than merely thought of or representative of something elsewhere, as we take ideas and beliefs to be.

Hume thinks that all impressions simply *seem* or *feel* immediately real, as ideas or beliefs do not, although he does allow a 'near resemblance in a few instances' (T 1.1.1.1, SBN 2). Consider hearing a distant sound, but then being uncertain as to whether you really heard it, or whether you imagined it: 'Was that really a sound', you might ask, 'or just the idea of a sound?' However, Hume's considered view is that it is only when one's mind is 'disordered by disease or madness' that one literally cannot distinguish impressions from

[5] Hume initially includes the term 'violence' among this variety of terms (T 1.1.1.1, SBN 1). In Book 2, however, where Hume distinguishes calm from violent passions, 'violent' means something like emotionally turbulent or intense (T 2.1.1.3, SBN 276). Along with most readers of Hume, I take the violence of a passion to be distinct from its vivacity: as an impression, any passion will be maximally vivid, but it may or may not be violent. However, see Radcliffe (2015b, 556) for an argument that the violence of a passion just *is* its vivacity or liveliness.

ideas, as one experiences them (E 2.1, SBN 17). If one's mind is not disordered, and if one is paying attention, then one cannot doubt that a sound that one hears is 'present to the mind' (T 1.1.7.4, SBN 19). The sound *just is* the mental object that one directly experiences, just as 'orange', 'sweet', and 'bitter' are 'objects'. (T 1.1.1.8, SBN 5). As Noonan (1999, 56) says, 'Hume reifies perceptions'. Any impression or idea in one's mind is a real object, directly present to consciousness.

Indeed, Hume appears more certain of the existence of impressions than he is of the existence of the world beyond our senses. He clearly distinguishes our sense impressions from the beliefs about any objects of 'real existence' which exist, as it were, 'behind' the impressions (T 1.4.2.24, SBN 199). These beliefs can be false, but impressions are not truth-apt in this way. Even if I think that an impression is illusory, as where I doubt that something that appears red is really red, I am certain that I experience the impression of redness. Hume thinks that one *cannot* doubt that any impression is a real existent, immediately present to one's mind. In Chapter 5, we will see that this has important implications for his theory of moral judgements.

Hume frequently defines perceptions by their level of vivacity, on a continuum from the liveliest, which are all impressions, to imagined, 'perfect' ideas, which are devoid of any vivacity (T 1.1.3.1, SBN 8). Ideas with some level of vivacity are typically either beliefs or memories: *a* belief is, for Hume, a vivid or believed idea. However, we will see in Chapter 3 that Hume also allows for a third category of vivid idea: some non-memory ideas become vivid, but without becoming beliefs. For want of a Humean name, I will call these 'quasi-beliefs'.

Hume argues that ideas can be derived from impressions in only two ways: either as memories or as perfect ideas. These are distinguished primarily by degrees in vivacity. An idea is a memory idea if it *immediately and directly* retains some of the vivacity of an

impression, so that it is 'somewhat intermediate betwixt an impression and an idea' in its level of vivacity (T 1.1.3.1, SBN 8). Presumably, Hume means it is intermediate between an impression and an imagined idea.

According to Hume, the primary distinction between memory ideas and those of the imagination is due to the greater vivacity of the former. However, he also notes a secondary distinction between the faculty of the memory and that of the imagination: the faculty of memory 'preserves the original form, in which its objects were presented, and… where-ever we depart from it in recollecting any thing, it proceeds from some defect or imperfection in that faculty' (T 1.1.3.3, SBN 9).

All non-memory ideas are formed as perfect ideas, which are entirely lacking in vivacity, but which we can manipulate in ways that we cannot do with memory ideas. One such way is by reasoning with them. Hume argues that there are only two forms of reasoning, both of which are performed by the faculty of the imagination. The first he calls 'demonstration', which consists of reasoning with 'the abstract relations of our ideas', as with mathematical reasoning (T 2.3.3.2, SBN 413). This form of reasoning has very little bearing on moral judgement, according to Hume, and neither the details nor the various interpretations of his account of it need concern us here.

The second form of reasoning consists of all 'reasonings from causation, and concerning matters of fact' (T 1.3.7.3, SBN 95). When Hume talks of beliefs, he officially refers only to the believed conclusions of causal reasoning.[6] Hume defines a belief, or believed idea, as a 'lively idea related to a present impression' (T 1.3.8.1, SBN 98). However, this is shorthand for lively (i.e. vivid) ideas related to impressions *or* memories (T 1.3.5.7, SBN 86). Indeed, Garrett (2015, 43) observes that Hume talks about an 'impression of the

[6] For a detailed argument to this effect, see Owen 2002, chapter 7.

memory', when he discusses their role in causal reasoning at T 1.3.4.1 (SBN 83), presumably because, like impressions, they are sufficiently vivid to begin a process of transferring vivacity to other ideas.

Causal reasoning is, as we shall see, precisely the process by which perfect ideas come to be vivid, via chains of associated ideas that connect them to our current impressions or memories. As Owen (2002, 157) puts it, ideas 'become beliefs by becoming more like impressions': as they acquire vivacity, we come to feel that they are ideas of real existents.

Causal reasoning is the product of two mental processes which are, for Hume, obviously prevalent in human psychology but ultimately inexplicable. These are the *association of ideas* and the *transference of vivacity between perceptions*. The first of these processes is described as the operation of 'a gentle force, which commonly prevails' whenever ideas are related by resemblance, contiguity in time or place, or cause and effect (T 1.1.4.1, SBN 10). Causal relations occur wherever we have frequently seen one object follow another, so that, purely from our experience of the 'repetition' of the one type of object following another, we come to associate the two ideas by custom or habit (T 1.3.8.10, SBN 102-3). For example, if we have often seen smoke follow fire, then any *new* perception of fire – whether an impression or idea – will be followed by an idea of smoke. Ideas of fire follow perceptions of smoke too, as we can make inferences both from cause to effect and from effect to cause. The stronger the association, the more swiftly and easily the second idea occurs.

The relevant kind of associations of ideas for causal reasoning is thus habitual or customary associations of ideas. This is a 'brute psychological principle, rather than a "principle" to which we consciously appeal in our causal reasoning' (Beebee 2006, 60). No person of sufficient experience – indeed, no animal of sufficient experience – can fail to form the idea of smoke when they experience a perception of fire, according to Hume.

However, even where ideas are strongly associated, this alone will not cause us to form beliefs. The second mental process required is the concurrent transference of vivacity from an impression or memory to an associated idea. If I see smoke rising over a distant hill, then the association of ideas between smoke and fire will impel me to form an idea of fire behind the hill. For me to *believe* this idea of fire, and thus to have reasoned that the smoke was caused by an unperceived fire, some of the vivacity of the perception of smoke must be transferred to the idea of fire. This is something which will naturally occur when the initial perception is an impression, so I will readily believe in the fire as a cause of the smoke, and as existing beyond my senses. In this way, causal reasoning is explained entirely by customary associations of ideas and the transference of vivacity. My present impression supplies the feeling of belief, and custom produces the idea that is to be believed.

It is thus from the operations of custom that we form beliefs about the world around us, according to Hume, and so build up a picture of a world stretching beyond our own perceptions, into the past, the future, and around us in space. In Chapter 5, I will ask whether he can allow any place for *moral* beliefs within this picture. First, however, we must understand the passion of approbation, along with its sympathetic causes. To this end, I will next consider Hume's theories of the passions and of sympathy. Hume understands sympathy as a psychological mechanism by which we may come to feel the passions of others, via a process that typically involves causal reasoning.

### 1.3. Passions and sympathy

Hume contrasts the passions, as 'secondary' or 'reflective' impressions, with the 'original' impressions of sense perception and bodily sensation, which appear to us 'without any antecedent perception' (T 2.1.1.1, SBN 275). By this, Hume means that original impressions have no observable psychological causes, whereas secondary impressions are caused by prior

perceptions. All secondary impressions 'proceed from some of [the] original ones, either immediately or by the interposition of its idea' (T 2.1.1.1, SBN 275).

Unlike ideas, secondary impressions do not always resemble the impressions that cause them, and they are not copies of the impressions that cause them. A headache or the sight of an ocean is an original impression, whereas an aversion to one's painful headache or a feeling of pleasure at the sight of an ocean is a secondary impression. All pleasant secondary impressions and desires are caused by pleasant impressions or ideas, whereas all unpleasant secondary impressions and aversions are caused by unpleasant impressions or ideas (T 1.1.2.1, SBN 7-8). All passions, therefore, including love and approbation, have determinable psychological causes.

Hume is sometimes thought to distinguish love from approbation by claiming that, although we love persons, we approve of 'characters, character traits, and motives' (Carlson 2014, 74; see also, e.g., Baier 1991, 134-5). In support of this claim, Carlson cites Hume's statement that, when we see someone who helps others, 'we approve of his character and love his person' (T 3.3.3.2, SBN 602). However, Hume does not appear to make any clear distinction between evaluating a person and evaluating one or more of that person's character traits. Consider the passage where he argues that we never approve of actions alone, but that 'the ultimate object of our praise and approbation is the motive, that produc'd them' (T 3.2.1.2, SBN 477). This certainly seems to suggest that we morally evaluate motives, rather than persons. Yet, throughout this section, Hume repeatedly mentions moral sentiments that *are* directed towards people: we 'blame a person' for not performing some action (T 3.2.1.3, SBN 477); where we find that someone has a virtuous motive, we feel 'esteem for him' (T 3.2.1.3, SBN 478); we 'blame a father for neglecting his child' (T 3.2.1.5, SBN 478).

Clearly, Hume emphasises that moral judgements are directed towards characters, character traits, and motives, rather than actions. However, he seems to assume, here and

elsewhere, that approbation is directed towards a person wherever it is directed towards one or more of her traits. Indeed, Hume often talks of a person's 'character' when he is considering a character trait *as* an evaluable part of the person whose trait it is (e.g. T 3.3.1.5, SBN 575; T 3.3.1.19, SBN 584; T 3.3.1.30, SBN 591). Unfortunately, this makes it harder to understand his distinction between love and approbation than Carlson suggests. We may approve of someone for her kindness, just as we may love her for her kindness.

To understand Hume's distinction between love and approbation, we must consider his theory of sympathy. Hume believes that sympathy may or may not be involved in our coming to feel love for someone, but that approbation occurs *only* where an idea of someone's character trait pleases us by sympathy (T 3.3.1.12, SBN 580).

Hume describes sympathy as a propensity 'to receive by communication [another person's] inclinations and sentiments, however different from, or even contrary to our own' (T 2.1.11.2, SBN 316). He claims that we can sympathise with people's beliefs, so that we may come to believe something simply because those around us do (T 2.1.11.2, SBN 316). More importantly for our purposes, however, he claims that we may sympathise with people's passions, pains, and pleasures.

Hume claims that 'sympathy is exactly correspondent to the operations of our understanding; and even contains something more surprising and extraordinary' (T 2.1.11.8, SBN 320). He argues that, wherever we are closely related to someone, our ideas of their passions are liable to become more lively than typical beliefs (T 2.1.11.3, SBN 317). The most relevant kind of relation here is resemblance. Hume claims that 'the idea, or rather impression of ourselves is always intimately present with us', and that we have as 'lively a conception of our own person' as we could possibly have of anything (T 2.1.11.4, SBN 317). Any idea of another person will show a 'great resemblance' to one's idea of oneself (T

2.1.11.5, SBN 318).[7] Any idea of that person's passion is thus likely to become livelier than a typical belief.

We saw, in §1.2, that Hume claims that ideas differ from impressions only in their levels of liveliness. Hume now reminds us of this claim: 'all ideas are borrow'd from impressions, and… these two kinds of perceptions differ only in the degrees of force and vivacity, with which they strike upon the soul' (T 2.1.11.7, SBN 318-9). If the 'component parts of ideas and impressions are precisely alike', then any sufficiently vivid idea would become an impression (T 2.1.11.7, SBN 319). And this is, indeed, what happens during sympathy, according to Hume: 'a lively idea is converted into an impression' (T 2.1.11.7, SBN 319). My idea of a person's happiness becomes so lively that I not only believe that she is happy; I *feel* happy.[8]

Our sympathies vary – we sympathise more with those who more closely resemble us than with others, for example – but Hume allows that we may sympathise with anyone, at least to some extent, simply because of our resemblance to one another as persons.

Just as Hume appears to see no important distinction between approving of a person and approving of that person's character trait, so he appears to see no important distinction

---

[7] Here, I will not consider how Hume understands the idea of the self in Book 2 of the *Treatise*. It is sufficient that he clearly relies on the existence of this idea.

[8] I agree with Stroud (1977, 198) that Hume only requires, and only really adheres to, a more general thesis of sympathy than he officially allows: 'Unpleasant feelings in others cause unpleasant feelings in us, and pleasant feelings cause pleasant feelings in us'. However, I do not see that this greatly influences Hume's discussion of approbation which, I will argue, is a pleasure *caused* by a sympathetic pleasure. This contradicts Abramson's (1999, 343) interpretation, by which our ideas of other people's pleasures and pains may be directly 'transformed' into moral sentiments via sympathy. The direct textual evidence regarding this point is unclear, for reasons that I will discuss in Chapter 6.

between sympathising with a person and sympathising with that person's passion. He claims both that we 'sympathize with others' (T 2.1.11.2, SBN 316-7) and that we 'sympathize with the passions and sentiments of others' (T 2.1.11.8, SBN 319). Moreover, the process which is specific to sympathy – its 'surprising and extraordinary' aspect – only occurs *after* we have formed an idea of a passion or sentiment, with which we then sympathise. This allows for Hume to claim that, once a relevant idea is in our minds, we may sympathise with it, even in some cases where it is not a typical belief about someone else's passion or sentiment. We will see the importance of this point in Chapter 3, when we consider Hume's thesis of 'delicate sympathy' (T 3.3.1.8, SBN 577).

For now, we can note the following. Given his claim that approbation is at least typically caused via sympathy with characters that have a 'tendency to the good of mankind', Hume seems to suggest that, if I see one person please another, then I will sympathise with the second person's happiness, and so approve of the first person (T 3.3.1.14, SBN 580). However, if my sympathy with the second person's happiness is the sole cause of my approbation, then it is very hard to understand Hume's theory of what we would now call the 'intentionality' of approbation: of the way that approbation can be *of* or *about* something or someone. As Cohon (2008, 175) argues, there appears to be a 'significant gap' in Hume's theory. He seems unable to explain how approbation may be caused via sympathy with one person, but then take another person as its intentional object.

To understand how approbation may be caused by sympathy, according to Hume, we must ask how he understands the intentionality of passions. And to be clear on this point, we must also ask how Hume understands ideas to represent *their* objects. I turn to these questions now.

### 1.4. Hume on intentionality and representation

Hume offers nothing like an explicit theory of the intentionality of the passions. Baier (1991, 160) thinks that, in one passage, Hume denies that passions *are* intentional, by claiming that a passion has no 'representative quality, which renders it a copy of any other existence or modification' (T 2.3.3.5, SBN 415). However, as Cohon and Owen (1997), Garrett (2006), and Merivale (2019) all argue, Hume merely denies that passions represent anything in the way that ideas represent impressions; by being copies of them. Indeed, we will see that he does not allow that *any* impressions can represent in this way.

Hume does not deny that impressions can represent external objects in *some* way. As Garrett (2015, 71) notes, Hume sometimes, albeit very rarely, suggests that sensory impressions represent external objects, as at T 1.2.1.5 (SBN 28). He also claims that we must take it 'for granted in all our reasonings' that there is a world beyond our senses (T 1.4.2.1, SBN 187). However, Hume does not mean by this that our sense impressions appear to us to be representative of anything beyond them. Unlike ideas, impressions do not appear, on introspection, to have *any* 'representative quality' (T 2.3.3.5, SBN 415).

It is partly because our impressions do not appear to be representative that Hume sees an irreconcilable problem in our understanding of the external world. In T 1.4.2, he argues that we firmly believe that there is an external world which we perceive via our senses, that we also believe that the objects of our senses may remain consistent even as our perceptions change, and that we have no sense of our perceptions as representative of any objects beyond them. The only way we can resolve these contradictions is, not via any philosophy, but by remaining careless and inattentive to them and simply assuming that there is 'both an external and internal world' (T 1.4.2.57, SBN 218).

Throughout all his arguments, Hume takes the following claim to be 'certain':

> [A]lmost all mankind, and even philosophers themselves, for the greatest part of their lives, take their perceptions to be their only objects, and suppose, that the very being, which is intimately present to the mind, is the real body or material existence (T 1.4.2.38, SBN 206).

Given this, Hume has no understanding of how impressions might represent objects, and he doubts that any such understanding is possible.

We have seen that Hume's Copy Principle entails that all ideas exactly resemble and are caused by simple impressions, and that they represent these impressions. Landy (2012) provides a helpful and, I think, highly plausible account of Hume's implicit 'semantic copy principle': that simple ideas represent impressions *by* being copies of them. Without going into the details of his argument, Landy (2012, 43) claims that Hume adheres to the following definition of 'representational content', at least insofar as it pertains to the representational content of simple ideas: '*x* has *y* as its representational content just in case *x* exactly resembles and is caused by *y*.'

Certainly, nothing that Hume says suggests that he sees how a perception could represent anything, unless it is both caused by that thing and exactly resembles it. This is part of his worry about our beliefs about the external world. If an impression represents something it must resemble that thing, and Hume cannot see how an impression can resemble anything other than a perception. The *only* form of representation that we can imagine obtaining for perceptions is that by which ideas represent impressions: by being copies of them (T 1.4.2.54, SBN 216). We derive our idea of a perception's capacity to represent only from our experience of ideas representing impressions by being copies of them. Therefore, given the Copy Principle, we can have no idea of any other way in which a perception can possess the capacity to represent.

As an important aside, I strongly doubt that Hume can consistently hold that impressions and ideas differ *only* in their degrees of vivacity, because he treats ideas as fundamentally representative, in a way that he thinks impressions cannot be. I am therefore sympathetic to Landy's suggestion that Hume implicitly sees impressions and ideas as distinct kinds of perceptions:

> Impressions are the original objects of the mind, derived from sources unknown; they are not copies of any other mental entities. Ideas are copies, either of impressions or of other ideas. It is this difference that makes a perception either an impression or an idea (Landy 2006, 124-125).

Landy (2006, 135) argues that, when Hume says that impressions and ideas differ from each other *only* in their different degrees of force and vivacity, he means that this is their only difference when 'considered individually, or non-relationally'. However, it is hard to reconcile even this claim with Hume's belief that our ideas represent as our impressions do not. Any impression is real simply in the sense that we are 'assur'd of its present existence', just as we are assured that our ideas exist *as* ideas (T 1.3.8.15, SBN 106). Unlike impressions, any idea is a real existent that is *also* representative of something that may or may not be a real existent, beyond our perceptions. Whereas an impression of redness just is the redness that one experiences when one sees red, an idea of redness is a perception which represents an impression of redness. This appears to be an important difference between the two perception kinds which is not *merely* a difference in vivacity. It is also a difference which may be observed even when impressions and ideas are considered individually and non-relationally.

In short, Hume's theory of ideas requires that all ideas represent as no impressions do: by being copies of prior perceptions. It is very hard to understand how, or whether, he can reconcile this with his claim that impressions and ideas fundamentally differ only in their different degrees of vivacity. However, I will not pursue this point here, but simply stress that Hume sees no other way that a perception could *be* representative, except by being a copy.

As passions are impressions, therefore, Hume needs to account for the way that a passion can be *about* its object, but without allowing that it represents its object. For this kind of reason, it is generally believed that Hume implicitly argues for an extrinsic explanation of the intentionality of passions, such that a passion can be of or about something only in virtue of its relation to other perceptions or objects.[9] For example, Cohon (2008, 165) claims that 'the intentional object of a passion seems… to be what the passion causes us to think of or attend to'. Similarly, Merivale (2019, 133) argues that Hume's 'default' position in the *Treatise* is that 'the object of a passion… is simply its cause'. I think both claims are correct, so that a passion is caused by that idea which it then causes us to attend to. Any idea which causes a passion is that idea which the passion is then *of* or *about*. I am sympathetic to Alanen's (2006, 194) view that Humean passions can be *of* objects by 'affecting the perception of whatever object they are reactions to'. My idea of someone who slighted me can cause a passion of anger, which then makes the idea of that person an unpleasant one to contemplate, in a way that is recognisably due to my anger.

Merivale (2019, 133) claims that, although most passions in Hume's *Treatise* are directed towards the ideas which cause them, there are 'a few notable exceptions, namely the passions of pride, humility, love, and hatred'. We will see, in Chapter 2, that Hume sees some

[9] Not everyone agrees. For example, Qu (2012) argues that Hume at least has the resources for an intrinsic account.

important parallels between these four passions. He devotes a significant proportion of Book 2 to developing a theory of their causes and objects, mainly via a discussion of pride. Several readers of Hume, including Árdal (1989), Davidson (1976), Korsgaard (1999), Merivale (2019), and Penelhum (1975), believe that his theory of pride entails the clearly implausible claim that there is only a contingent relationship between pride and the idea of oneself, such that we feel proud *before* we think of ourselves, and so feel proud of ourselves. In Chapter 2, I will argue, against this general interpretation, that Hume applies his default, causal theory of intentionality to pride, humility, love and hatred, as well as to all other passions.

It is, of course, far from obvious whether anything like Hume's causal theory of intentionality could be a successful one. To address one important worry, he believes that 'to form the idea of an object, and to form an idea simply is the same thing' (T 1.1.7.6, SBN 20). For Hume, our experiences just are our perceptions; we can only conceive of something by forming an idea of it; and we can only form ideas of that which we have experienced. Hume concludes from this that we cannot possibly conceive of anything other than our perceptions. I agree with Cohon and Owen (1997, 55) that this is his point in the following passage:[10]

> [S]ince nothing is ever present to the mind but perceptions, and since all ideas are deriv'd from something antecedently present to the mind; it follows, that 'tis impossible for us so much as to conceive or form an idea of any thing specifically different from ideas and impressions (T 1.2.6.8, SBN 67).

[10] Garrett (2006, 306) argues that Hume is making a somewhat subtler point in the passage.

If we cannot conceive of anything except by forming an idea of that thing, then this suggests that, wherever we think of any person or thing in the world, we can make no meaningful distinction between that person or thing and our idea of that person or thing. I think this is indeed Hume's view. If so, then we also cannot make any meaningful distinction between feeling a passion towards a person or thing in the world and feeling that passion towards our idea of that person or thing.

Hume certainly seems to assume precisely this about the passions of pride and humility. He sometimes claims that an *idea* 'excites' pride or humility (T 2.1.5.5, SBN 286), or that an *idea* is that 'to which [pride and humility] direct their view, when excited' (T 2.1.2.4, SBN 278). At other times, and without any suggestion that he is providing an alternative view, he describes both the causes and intentional objects of pride as the *objects* of ideas: 'the cause [of an instance of pride] is the beautiful house' (T 2.1.2.6, SBN 279); we are proud of 'power, riches, beauty or personal merit' (T 2.1.3.4, SBN 281). Cohon (2008, 165) and Penelhum (1975, 99) both note Hume's ambiguous language in this regard.

We will see in §2.1 that Hume thinks that pride and humility are importantly similar to love and hatred. This strongly suggests that he sees no meaningful distinction between loving or approving of a person and loving or approving of the idea of that person. Regardless of any worries we might have about this claim, I will argue, in Chapters 2 and 3, that it at least allows Hume to apply his default, causal theory of intentionality to pride, humility, love and hatred, and to the moral sentiments.

In Chapter 2, I will address Hume's theory of the passions in greater detail, and I will discuss two important aspects of this theory. The first is his distinction between direct passions and indirect passions. The second is his distinction between calm passions, such as approbation, and violent passions, such as love. In Chapter 3, I will conclude from this discussion that Hume understands love as a violent sentiment, which may or may not be

caused via sympathy, and that he understands approbation as a calm sentiment of the same general kind as love, but caused in all cases via a ‘delicate’ kind of sympathy (T 3.3.1.8, SBN 577). We will see that approbation is always caused by, and directed towards, a complex idea that is, roughly, of a person with a trait of some kind that is such as to cause happiness.

# 2. Hume on Indirect and Calm Passions

In this chapter, I address those aspects of Hume's theory of the passions that are most relevant to his theory of moral judgements, with a focus on his distinctions between direct and indirect passions, and between calm and violent passions. I begin this chapter by arguing that Hume sees a generally unappreciated but vital role for the formation of complex ideas in the production of four indirect passions: pride, humility, love, and hatred. This allows him to develop an account of their intentionality that is more coherent than many think possible.

I then argue for a new interpretation of Hume's distinction between calm and violent passions, according to which calm passions are calm because they are caused by more generalised ideas than those that cause violent passions. This allows Hume to distinguish love from approbation in a way that has hitherto been unrecognised. We will see, in Chapter 3, that he understands love as the violent passion by which we come to value people for their particular pleasing traits, attributes, or possessions. He understands approbation as the calm passion by which we come to value people who possess token character traits of generally pleasing types.

Mine is by no means the only interpretation of Hume's account of the passions, of course: Fieser (1992) and Radcliffe (2015a) provide useful surveys of the many different ways that Hume's theory and taxonomy of the passions may be understood. I will address rival interpretations to my own at key points, but I cannot hope to give detailed arguments against each one. Instead, I will argue for my own interpretation, after which the proof will be in the pudding. In Chapters 3 to 6, I will consider several aspects of Hume's theory of moral judgements which, I believe, provide strong support to my arguments within this chapter.

In §2.1, I address Hume's account of the passions of love, hatred, pride, and humility, and I argue that he believes that these are each caused by certain kinds of complex ideas. In §2.2, I argue that Hume understands violent passions to be caused by particular ideas, and

calm passions to be caused by general ideas. Approbation is a *calm* form of love, so I will conclude that the complex idea that causes approbation must be or include a general idea.

### 2.1. Indirect passions

Hume gives the following account of the direct/indirect distinction: 'By direct passions I understand such as arise immediately from good or evil, from pain or pleasure. By indirect such as proceed from the same principles, but by the conjunction of other qualities' (T 2.1.1.4, SBN 276). The indirect passions include 'pride, humility, ambition, vanity, love, hatred, envy, pity, malice, generosity, with their dependants' (T 2.1.1.4, SBN 276-7).

Hume, rather unhelpfully, sometimes uses the term 'passion' to refer to all impressions of reflection, and sometimes to refer only to violent passions. Consider the claim that, when we 'take a survey of the passions, there occurs a division of them into *direct* and *indirect*' (T 2.1.1.4, SBN 276). Loeb (1977, 396) understands Hume to mean by this that only *violent* passions can be direct or indirect. However, approbation is always calm. Loeb concludes that approbation is neither direct nor indirect, so that it cannot be a form of love. Conversely, if approbation is a calm form of *love*, as Hume claims, then it must be an indirect passion.

Love is one of four indirect passions, along with hatred, pride, and humility, that are caused via a 'double relation of ideas and impressions' (T 2.1.5.5, SBN 286). Merivale (2019, 133) helpfully calls these the 'double-relation passions'. As Árdal (1989), Cohon (2008), and Taylor (2015) argue, at least some indirect passions are evaluative passions, in a way that desires, for example, are not. We generally desire what we already value. In contrast, to use Cohon's (2008, 172) helpful phrase, the double-relation passions 'do not so much track value as impart it'. Love is that passion by which we come to value others, just as pride is the passion by which we come to value ourselves. For Hume, the good just is the

pleasing (e.g. T 2.1.1.4, SBN 276). Therefore, all double-relation passions must be pleasures or pains.

Radcliffe (2004, 650) notes that Hume has an 'unofficial line' in which he treats passions 'as though they are defined in terms of the processes that produce them'.[1] Hume claims that a passion like love is indefinable, and can only be known via experience (T 2.2.1.1, SBN 329). Nevertheless, he allows that each distinct, indefinable passion has observable and consistent causes, and he frequently defines passions, tacitly or otherwise, in terms of these causes. For example, pride is 'that agreeable impression, which arises in the mind, when the view either of our virtue, beauty, riches or power makes us satisfy'd with ourselves' (T 2.1.7.8, SBN 297).

I will argue that, once we understand their causes, we will see that love and approbation have different, but closely related, objects. We love a kind person for her particular motive of kindness, which may please in a variety of ways, self-interested or otherwise. We approve of a kind person simply for having a motive of a type that we generally associate with pleasing others. Each passion is caused by, and directed towards, a complex idea that includes an idea of the person and an idea of her motive.

This may seem implausible at first glance, because Hume argues for 'a distinction betwixt the cause and the object' of each of the double-relation passions (T 2.1.2.4, SBN 278). However, I will argue that Hume merely distinguishes the *distal* cause of each double-relation passion from its object. What Hume calls the 'cause' of a double-relation passion is the idea that 'gives the first motion' to that passion, by initiating the process that causes it (T 2.1.5.8, SBN 288). The *proximate* cause, and object, of any double-relation passion is a

[1] Davidson (1976, 754) holds a similar view.

complex idea of some person, considered as the possessor of some pleasing or displeasing object.[2]

Hume's discussion focusses mainly on pride. He argues that pride cannot be caused simply by the idea of the self, because both pride and humility are felt towards the self. If the idea of the self directly caused these passions, then they would both consistently occur, so that they would, in effect, cancel each other out (T 2.1.5.7, SBN 288). Therefore, Hume claims, a different idea must cause pride. Any such idea must be of something that causes pleasure, such as a beautiful house. As I interpret him, he argues that, wherever one is closely related to anything by causation or contiguity, the idea of that thing will bring the associated idea of oneself to mind, and the two ideas will combine to form a more complex idea. If the idea of that which is related to oneself is a pleasing idea, then the more complex idea will cause pride. The proximate cause and object of pride is thus always a complex idea that contains the idea of the self, such as that of *myself as the owner of a beautiful house*.

To argue for this interpretation, I will first distinguish the causal features of a double-relation passion into two, each of which is a necessary condition (and both of which are jointly sufficient) for that passion to meet the definition of a double-relation passion.

The first necessary condition for any passion to be a double-relation passion is that it is felt towards something that already causes a different pleasure (for love or pride) or pain (for hatred or humility). Hume claims that at least two 'resembling impressions are connected together' whenever a double-relation passion occurs, one of which will be the double-relation passion itself (T 2.1.4.3, SBN 283). If I am proud of a house that is aesthetically pleasing, then this aesthetic sentiment resembles pride by being a pleasant feeling, as pride is.

[2] Schmitter (2008, 231) suggests something like this possibility. However, her considered view is that the intentionality of pride is due to a network of perceptions, which together draw our attention to our 'character, or personality' (Schmitter 2008, 236).

The second necessary condition for any passion to be a double-relation passion is that a relevant association of ideas occurs in its production, such that the idea of some pleasing or painful object is associated with the idea of the self or of someone else. These ideas may be related by 'contiguity', but they are more typically related by 'causation' (T 2.1.9.4, SBN 305). Anything sufficiently closely related to oneself by contiguity or causation may cause pride, from a fine house to a beautiful face, a virtuous motive, power, wealth, or a beautiful country.

At some points, Hume claims that, when some relevant object pleases us, we feel a passion of pride, which *then* causes us to think of the self. For example, he describes pride as a 'passion plac'd betwixt two ideas, of which the one produces it, and the other is produc'd by it' (T 2.1.2.4, SBN 278). It is clearly the *second* idea which Hume sees as that of the self.

Given phrases like this, it is easy to see why Árdal (1989, 388) believes that Hume misrepresents the relation between pride and the self 'as a causal relation between pride and the thought of oneself. Whenever one is proud, one's thought is drawn to oneself, but according to Hume it could have been otherwise'. Call this putative claim the 'contingency thesis': that feeling proud causes one to think of the self, and so to then feel proud *of* oneself. Hume is often believed to argue for this thesis (e.g., Davidson 1976; Korsgaard 1999; Merivale 2019; Penelhum 1975). Clearly, however, it cannot be right: pride *just is* a form of self-evaluation, and so the relation between pride and the self cannot be contingent in this way.

Hume undoubtedly suggests the contingency thesis, more than once. However, despite the weight of evidence, he surely cannot mean to endorse it. Schmitter (2008, 233) briefly notes that, because the idea of one's possession is 'linked by relations of association' to the idea of oneself, Hume must believe that the former idea will immediately produce the latter idea, *without* requiring the presence of pride. But this point is crucial: Hume would be

very unlikely to claim that *any* identifiable perception is involved in the process by which the idea of one's property brings the idea of oneself to mind. The principle of the 'association of ideas' is integral to Hume's philosophy, and he clearly intends a central role for it in his theory of pride (T 2.1.4.2, SBN 283). Yet, as Davidson (1976, 749) observes, the contingency thesis denies it any such role.

Hume believes that we are typically proud of things that are *causally* related to ourselves. And, according to Hume, 'there is no relation, which produces a stronger connexion in the fancy, and makes one idea more readily recall another, than the relation of cause and effect betwixt their objects' (T 1.1.4.2, SBN 11). Why, then, would pride play *any* role in producing the idea of the self, once one forms the associated idea of one's property, virtue, wealth, or power? Hume would find it difficult, to say the least, to distinguish any such role from that of the association of ideas.

In fact, after T 2.1.5.6 (SBN 286-7), Hume no longer suggests that the feeling of pride produces an idea of the self. He starts claiming instead that the idea of the self is *involved* in the cause of pride. Merivale (2019, 135) believes that he simply equivocates over 'whether the object of the double-relation passions is their effect or a part of their cause'. However, I think this point marks a rhetorical shift in Hume's (admittedly confusing) argumentative strategy.

### *2.1.1. The associative causes of the double-relation passions*

I agree with Cohon's (2008, 166) assessment, that Hume wants to explain how a 'shift in attention' can occur when we feel any double-relation passion, such that thinking about an object will make us think of a person. Hume believes that this always happens with pride and humility, as where thinking of a house makes me think of, and feel proud of, myself as its owner. I also agree with Cohon (2008, 162) that Hume 'uses his associationism to explain the

generation' of these passions. Unlike Cohon, however, I think that Hume *only* suggests the contingency thesis to stress that a shift in attention always occurs when we come to feel pride or humility, so as introduce the need for his associationism.

Kemp Smith (1966, 166) suggests that the length of Hume's treatment of the double-relation passions is 'significant as showing his preoccupation with, and sense of, the revolutionary character of his teaching in regard to association at the time when this part of the *Treatise* was being written'. I suggest, moreover, that Hume's efforts to persuade his readers that *something* causes a shift in our thinking when we feel pride signifies his belief that many of his readers will as yet be unconvinced by his theory of association. By claiming that pride produces the idea of the self, Hume is, I think, merely laying the groundwork to argue that an association of ideas produces the idea of the self, during the process that causes pride. He wants to emphasise that pride is always felt towards the (idea of the) self, despite always being 'excited' by an idea that is not of the self, so that he may offer his own, associative explanation for this shift in attention (T 2.1.2.4, SBN 278).

Cohon (2008, 165) assumes, as Hume's readers generally do, that he means the same thing when he says that pride 'turns our view to ourselves', and that pride 'naturally produces' the idea of the self (T 2.1.5.6, SBN 287). However, it is not obvious that the meaning of these phrases is identical. Hume only suggests that pride produces the idea of the self *before* T 2.1.5.7 (SBN 287-8). Here, he has only just started arguing for the incompatible claim that pride occurs at the end of the process of a 'double relation of ideas and impressions', during which an idea with a 'relation to self' produces, by association, the idea of the self (T 2.1.5.5, SBN 286). Once he has claimed that this associative process *produces* the idea of the self, in T 2.1.5.8 (SBN 288), he no longer claims that pride does so. Indeed, after this, he only claims that pride 'turns our view' to the self on one further occasion: in a passage where he also claims that the sexual appetite turns our view to the idea of sex, which

it typically does *after* that idea has caused the sexual appetite (T 2.2.11.6, SBN 396). We should, therefore, ask if we can plausibly interpret Hume's phrase 'turning our view to an idea' to mean something other than 'producing an idea'.

I suggest one such interpretation: that for a passion to turn one's view to something is for it to cause us to continue to think about that thing, so long as the passion is present to the mind. Consider cases where one cannot stop thinking about one's beloved, or where a proud person continually thinks of, and talks about, himself. Similarly, Hume may mean that, once we start thinking about sex, we keep thinking about sex. So interpreted, the claim that pride turns one's view to oneself entails that, *once* one feels proud of oneself, one cannot 'ever lose sight of this object' so long as the feeling is present (T 2.1.5.3, SBN 286). However, it suggests nothing about how pride is caused, or about how it comes to take the self as its object.

Whether or not my suggestion is correct, Hume's theory of the double relations of ideas and impressions is, I will argue, incompatible with the contingency thesis. His theory entails that an 'original and natural instinct' ensures that the feeling of pride is always caused by, *and* directed towards, a complex idea of oneself as related to some pleasing object (T 2.1.5.3, SBN 286). Unfortunately, Hume discusses this instinct before he has explained his theory of the causes of pride. At this stage, all he has '*establish'd*' is that pride is pleasing, and that its object is the self (T 2.1.5.5, SBN 286). This is, I suggest, why he merely asserts that instinct makes pride take the idea of the self for its object, not for its cause.

Hume's last, clear mention of the contingency thesis is his claim that pride, 'after its production, naturally produces' the idea of the self, just as hunger naturally makes us think of food (T 2.1.5.6, SBN 287). At this point, Hume is arguing that pride, by instinct, 'always turns our view to ourselves, and makes us think of our own qualities and circumstances' (T 2.1.5.6, SBN 287). He may mean to assert the contingency thesis merely to introduce this

claim, or he may mean to suggest *only* that, just as we cannot stop thinking of food when we feel hungry, so we cannot stop thinking of ourselves and our qualities and circumstances when we feel proud. Either way, in the following paragraph, he asks anew what the causes of pride are. He answers that instinct alone cannot cause us to feel proud of ourselves, and that some 'foreign object' must therefore cause pride (T 2.1.5.7, SBN 287).

Hume only *then* claims to 'discover this cause, and find what it is that gives the first motion to pride' (T 2.1.5.8, SBN 288). He argues that every 'cause' of pride is 'ally'd to the object of the passion' or, in other words, is related to the self (T 2.1.5.8, SBN 288). Every 'cause' also causes a pleasing passion, to which pride will be related by resemblance. Hume concludes that these two relations constitute the 'very principle, which gives rise to pride' (T 2.1.5.8, SBN 288). After this, he no longer suggests that the feeling of pride produces the idea of the self, presumably because he has now argued that this idea is produced by an association of ideas instead. He now clearly states that a relation between some object and the self is required *before* pride can occur: 'Any thing, that gives a pleasant sensation, and is related to self, excites the passion of pride, which is also agreeable, and has self for its object' (T 2.1.5.8, SBN 288).

Once Hume has established that the idea of the self is not sufficient to cause pride, he claims that it is a *necessary* part of the cause: 'In order to excite pride, there are always two objects we must contemplate, *viz*. the *cause* or that object which produces pleasure; and self, which is the real object of the passion' (T 2.1.6.5, SBN 292). Merivale (2019, 135) sees this as Hume equivocating over the causal relation between the idea the self and pride, but I believe that he has by now firmly rejected the contingency thesis, which would otherwise violate his general rule that a 'cause must be prior to [its] effect' (T 1.3.15.4, SBN 173). Pride *cannot* produce the idea of the self, if that idea must be contemplated to produce pride.

Hume also now claims that pride has 'in a manner two objects, to which it directs our view' (T 2.1.6.5, SBN 292). Previously, he has insisted that the object of pride is the self. However, he has never mentioned being proud *of* oneself; only of one's possessions, such as 'beauty', (T 2.1.2.5, SBN 279), or a 'beautiful house', (*T* 2.1.2.6, SBN 279), or 'handsome chairs and tables' (T 2.1.3.5, SBN 281). He presumably means by these phrases that one may be proud of oneself *as* a beautiful person, or *as* the owner of a beautiful house, or *as* the owner of handsome chairs and tables. The object of pride is, therefore, 'in a manner' two objects although, in another manner, it is only one: it is a complex idea of the self as related to some pleasing thing.

Assuming that Hume ultimately rejects the contingency thesis, he can, entirely consistently, claim that this complex idea is also the proximate cause of pride. This would cohere with his default, causal theory of the intentionality of passions, and with his general rule that causes precede effects. This is, I suggest, the most charitable and most plausible interpretation of his argument.

Hume's considered view is that an association between an idea of an object and the related idea of the self is necessary to cause pride or humility: A 'relation of ideas' is 'requisite… to the production of the passion [of pride or humility]' (T 2.1.9.5, SBN 305). We feel pride 'upon the appearance of a related object', and so *after* a 'relation or transition of thought' has occurred (T 2.1.9.5, SBN 305). The only relevant transition of ideas is that between the idea of some pleasing or painful object and the idea of the self. This transition is needed to 'second a relation of affections, and facilitate the transition from one impression to another' (T 2.1.9.5, SBN 305). Hume means that pride or humility will be the *second* impression to occur: his point is that any pleasures or pains can only be relevant to the causes of pride or humility if we experience them before we feel pride or humility. Pride occurs only

at the end of the process by which a double relation of ideas and impressions causes it to occur.

To take stock: after T 2.1.5.6 (SBN 286-7), there are no obvious statements of the contingency thesis. Instead, Hume argues that, where any idea of a pleasing thing is related to the self, it will produce the idea of the self by association, before pride occurs. Throughout, he implies that the object of pride is a complex idea of oneself as related to a pleasing thing. This strongly suggests that the process that causes pride involves the two ideas combining, to form the relevant complex idea.

When Hume first discusses the associations of ideas, in Book 1, he is primarily interested in the 'associating quality' that causes simple ideas to 'fall' into complex ones (T 1.1.4.1, SBN 10). Admittedly, he never uses the phrase 'complex idea' during his account of pride. Nevertheless, when he first mentions the double relation of ideas and impressions, Hume claims that any idea that causes pride 'is easily converted into its cor-relative': the idea of the self (T 2.1.5.5, SBN 286). He cannot mean that the one idea is *literally* converted into the other, but he does allow that ideas 'are capable of forming a compound by their conjunction' (T 2.2.6.1, SBN 366). Any such compound can only be a complex idea. And the most likely complex idea to cause pride, from all that Hume says, is that of the self as related, by causation or contiguity, to some pleasing thing.

This interpretation is consistent with everything that Hume says, apart from the contingency thesis. We have seen very good reasons to think that he ultimately rejects this thesis. Absent this thesis, and given his default, causal theory of the intentionality of passions, Hume can treat it as a conceptual truth (because a matter of definition) that the passion called 'pride' is directed towards a complex idea of the self, as related to some pleasing object. Pride just is that pleasure which is caused by such ideas (T 2.1.7.8, SBN 297).

There is some support for this interpretation in Hume's later *Dissertation on the Passions*. Merivale (2019, 142) notes an interesting change between two otherwise near-identical claims that Hume makes in the *Treatise* and in the *Dissertation*. In the earlier work, he claims that an injured person may 'find a hundred subjects of discontent, impatience, fear, and other uneasy passions; especially if he can discover these subjects in or near the person, who was the *cause* of his first passion' (T 2.1.4.4, SBN 284, my emphasis). In the *Dissertation,* he says almost the same thing, except he concludes with the mention of 'the person, who was the *object* of his first emotion' (P 2.8, Bea 8, my emphasis).

Merivale (2019, 142) sees this as evidence that Hume may have altered his theory of pride between the two works, although he confesses that, aside from this one alteration, 'Hume's presentations of the double-relation theory in the Treatise and the Dissertation are almost exactly identical, differing mainly in superficial points of style'. Merivale thinks that, unless he was making a substantive change to his view, Hume could only have been correcting a mistaken use of the word 'cause' in the *Treatise*. However, I think a third possibility more likely than either of these: Hume's alteration was simply another superficial change, because he understands *both* the proximate cause and object of the person's hatred as a complex idea of some other person who possesses some displeasing characteristic.[3]

Hume devotes much more attention to pride and humility than to love and hatred. However, he argues that the 'same qualities that produce pride or humility, cause love or

[3] Merivale (2019, 142) describes the relevant passion as 'anger' rather than hatred. Hume does not name the passion in question, but he describes the person as 'very much discomposed and ruffled in his temper', which certainly supports this (P 2.8, Bea 8; see also T 2.1.4.4, SBN 284). However, both the *Treatise* and the *Dissertation* passages occur during his discussions of the double-relation passions, so it seems likely that he is discussing one of these. The only plausible contender here is hatred.

hatred', but where the idea of the self is replaced by the idea of some other person (T 2.2.1.9, SBN 332). There is, however, a further, important difference: pride and humility do not typically cause desires, whereas 'love and hatred are always followed by, or rather conjoin'd with benevolence and anger' (T 2.2.6.3, SBN 367). We want to see those we love happy, and those we hate unhappy.

Given the foregoing, we can conclude that Hume understands love as the kind of pleasure that is caused by a complex idea of some person other than oneself, as related, by causation or contiguity, to some pleasing thing. To give an example, if I see one person help another, then I am likely to form a complex idea of the first person as the cause of the second person's happiness. The idea of this happiness will please me, via sympathy. The more complex idea will therefore cause the resembling pleasure of love.

The instance of love just described is clearly similar to approbation, as Hume understands it. Unlike approbation, however, it will be a violent passion of love. Nevertheless, we can already see how we might resolve Cohon's worry, discussed in §1.3, regarding the seeming gap between cases where we sympathise with one person and so approve of another. Hume *has* explained, albeit indirectly, how we can love one person when we sympathise with another: because, in such cases, love is caused by a complex idea of the first person as the cause of the happiness of the second, where this happiness pleases us by sympathy. The explanation concerning approbation must be a similar one.

In Chapter 3, I will argue that Hume claims that all ideas of character traits of generally pleasing types cause a certain, *calm* kind of sympathetic pleasure. Wherever we form a complex idea of a person as possessing such a trait, this complex idea will then be the cause and object of a similarly calm pleasure of approbation. Before considering this point in detail, however, we must consider Hume's account of the calm passions.

## 2.2. Calm passions

'Calm' passions are those that 'produce little emotion in the mind, and are more known by their effects than by the immediate feeling or sensation' (T 2.3.3.8, SBN 417; see also T 2.1.1.3, SBN 276). Baier (1991, 164) notes that Hume and his contemporaries generally mean by 'emotion' a 'bodily disturbance'. A calm passion is one without any noticeable degree of *mental* disturbance. A violent passion is one that presents to us as agitated, intensely felt, or in some other way mentally disturbing. Calm passions are ones that do not present to us in this way.

Hume does not suggest that we cannot observe the presence of calm passions by introspection: he does not claim that the passions *themselves* are 'nearly imperceptible' or – worse – 'imperceptible perceptions', as Schauber (1999, 362) suggests. For Hume, perceptions just are those objects that are 'immediately present to us by consciousness', and calm passions are perceptions (T 1.4.2.47, SBN 212). They are no less 'real passions' than the violent ones that we typically call 'passions' (T 2.3.3.8, SBN 417). As Paxman (2015, 273) argues, they must possess a noticeable, albeit 'subtle and tranquil', feeling. This causes us to frequently mistake them for the 'determinations of reason' (T 2.3.3.8, SBN 417).

Hume discusses calm passions at two key points in the *Treatise*. The first of these is at the very beginning of his book on the passions, when he briefly introduces the distinction between calm and violent passions (T 2.1.1.3, SBN 276). He claims, for example, that a typically calm pleasure felt towards music may become violent. Árdal (1966, 94-95) understands this to mean that 'a passion classified as calm can, upon occasion, be violent', as where we move from 'calmly enjoying something' that we typically calmly enjoy to being 'completely carried away'. I think this is correct, so far as it goes. However, I think Hume's point, although brief, is more important than just this.

Hume claims that what we typically call 'passions' are 'more violent than the emotions arising from beauty and deformity', so that we make a 'vulgar and specious division' between them (T 2.1.1.3, SBN 276). For example, fear is commonly called a 'passion', whereas aesthetic disapproval is not, but they are nevertheless both passions. Hume will argue for this in greater detail in Book 3, as we will see in Chapter 3.

Hume describes aesthetic approval as a calm 'emotion arising from beauty', by which he can only mean that it is caused by whatever is beautiful in virtue of its occurrence (T 2.1.1.3, SBN 276). This calm passion is contrasted with the 'raptures' of music: violent emotions, presumably caused by music that we find beautiful (T 2.1.1.3, SBN 276). When the music finishes, our violent pleasure 'may decay into so soft an emotion, as to become, in a manner, imperceptible' (T 2.1.1.3, SBN 276). Hume seems to be suggesting that our taste in music is a very calm form of the *same* kind of joy that listening to music may cause. Similarly, the 'sense of beauty and deformity in action' appears to be a calm 'kind' of a certain passion, of which the violent kind comprises the 'passions of love and hatred' (T 2.1.1.3, SBN 276). This already suggests that calm forms of love and hatred constitute what Hume will later (in T 3.1.2) call the 'moral sense': the sentiments of approbation and disapprobation.

The second key discussion of calm passions in Book 2 concerns only *direct* passions: 'certain calm desires and tendencies' (T 2.3.3.8, SBN 417). These are of two kinds: 'either certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children; or the general appetite to good, and aversion to evil, consider'd merely as such' (T 2.3.3.8, SBN 417). Hume argues that violent desires are typically more motivationally powerful than calm ones (T 2.3.8.13, SBN 437-8). However, he allows that calm desires can sometimes be 'strong', or powerfully motivating, and so

overcome more violent passions (T 2.3.4.1, SBN 418-9). Hume claims that this is what we call 'strength of mind' (T 2.3.3.10, SBN 418).

The passions that Hume calls 'desires' and 'aversions' at T 2.3.9.1 (SBN 438) are violent passions, caused by impressions or ideas of 'pain and pleasure'. Cohon (2008, 164) argues that Hume distinguishes between desires that are caused in this way and *instinctive* desires, such as hunger, lust, and the desire to be kind to children. I broadly agree, although I disagree with Cohon's (2008, 164) claim that instinctive desires occur in ways 'for which we cannot give any causal explanation'. Hume believes that *all* passions are caused by identifiable, prior perceptions: they are 'secondary' impressions (T 2.1.1.1, SBN 275).

Kemp Smith (1966, 168) argues that the instinctive passions are neither direct nor indirect: they are *not* desires, but what he calls 'primary' passions, which play the roles typically taken by pains and pleasure in the production of desires. However, this cannot be right: one such passion is clearly named as 'the desire of punishment to our enemies' (T 2.3.9.8, SBN 439). This is the violent form of that instinctive desire which, when calm, Hume calls 'resentment' (T 2.3.3.8, SBN 417).

Hume discusses several violent, instinctive desires within his account of the direct passions, in the paragraph immediately following his discussion of the violent desires and aversions that are caused by perceptions of pain or pleasure. The instinctive desires appear to be desires and aversions that are caused by perceptions *other* than ideas of any pleasure that will be gained, or of any pain that will be avoided, by acting on them. What Hume calls 'hunger' and 'lust' are, presumably, violent desires caused directly by physical sensations (T 2.3.9.8, SBN 439). Hume claims that some *calm* desires are also instinctive, or caused by perceptions other than ideas of pain or pleasure. The 'love of life', for example, is presumably the desire to avoid death, which will be caused by ideas of one's death; even, presumably, in cases where one believes in a pleasurable afterlife (T 2.3.3.8, SBN 417).

In fact, all passions are fundamentally instinctive: each type of passion is caused, seemingly as a matter of brute fact, by some identifiable type of perception or set of perceptions. Double-relation passions occur as they do only by instinct (T 2.1.5.3, SBN 286). Instinct ensures that benevolence is caused by the passion of love (T 2.2.6.6, SBN 368). It is also 'instinct' that ensures that perceptions of pleasure cause desires and that perceptions of pain cause aversions (T 2.3.9.2, SBN 438).

In summary, Hume appears to merely distinguish those desires that are caused by perceptions of pleasure or pain from those that are caused by other perceptions. Most desires, as discussed at T 2.3.9.2, are instinctively caused by feelings of pleasure (which one then desires to continue) or of pain (which one then desires to avoid), or by ideas of some pleasure that will occur if one acts, or of some pain that will occur if one does not act. Some desires, such as hunger, benevolence, and the desire to be kind to children, are at least sometimes instinctively caused by other kinds of perceptions than these. Nevertheless, for want of better terms, I will call the former kind 'learned' and the latter kind 'instinctive'.

With this in mind, we can see that Hume provides a taxonomy of three types of calm passion. In T 2.1.1.3 (SBN 276), he discusses *calm pleasures and pains*, which appear to include the moral passions: calm forms of the indirect passions of love and hatred. In T 2.3.3.8 (SBN 417), he discusses *calm instinctive desires*, such as the desires to be kind to children and to avoid death, and *calm learned desires*, which he summarises as 'the general appetite to good, and aversion to evil, consider'd merely as such'. Later, he discusses similar, violent desires. *Violent instinctive desires* include those for food and sex, as well as the 'desire of punishment to our enemies' (T 2.3.9.8, SBN 439). *Violent learned desires* are those that arise 'from good [and evil] consider'd simply' (T 2.3.9.7, SBN 439).

All three kinds of calm passion appear to be easily confused for reasoned beliefs. Moral sentiments, and other sentiments of taste, are often mistaken for evaluative beliefs.

For example, if I sincerely say that blues music is a valuable and important genre, then I am, it seems, expressing a calm pleasure at the thought of blues music. This is an expression of my taste, and taste is 'plainly nothing but a sensation of pleasure' (T 2.1.7.7, SBN 297). I am, of course, aware that I take this view: introspection reveals an evaluative perception. However, I am likely to mistake it for a belief about the value of blues music, because it feels far less violent than my feelings of pleasure when I listen to, for example, Nina Simone singing 'Nobody's Fault but Mine'. What I take to be a perception that represents the value of blues music is, in fact, a calm perception of pleasure, directed towards my general idea of blues music, and with no representative properties.

Calm instinctive desires seem unlike mere preferences: it seems *reasonable* to seek the best for our loved ones, or to be kind to children, or to avoid death. Calm learned desires presumably feel like beliefs about how we *ought* to act, as where we feel that we ought not perform an action because we disapprove of the idea of it.

Our question now is, how should we understand the *causes* of these passions?

### *2.2.1. What makes a calm passion calm?*

Loeb (2005, 4-6) argues that a 'calm' passion is one that forms a settled and stable element within one's psychology, whereas violent passions are 'volatile'. I do not think this is what Hume *means* by 'calmness' or 'violence'. In both T 2.1.1.3 (SBN 276) and T 2.3.3.8 (SBN 417), he quite clearly uses 'calmness' to refer to a lack of emotional feeling. However, I somewhat similarly understand Hume to argue that we experience calm passions more frequently than violent ones, and that they are consistently experienced towards the same kinds of objects, as violent passions are not. Certainly, Loeb (2005, 4) points to something interesting when he claims that one passage in particular is 'suggestive of a more fundamental distinction [between calm and violent passions] than emotional intensity'.

In the relevant passage, Hume argues that the calmness of a passion is not simply due to its motivational weakness, but rather that 'when a passion has once become a settled principle of action, and is the predominant inclination of the soul, it commonly produces no longer any sensible agitation' (T 2.3.4.1, SBN 418-9). Hume then makes a further claim that Loeb pays surprisingly little attention to:

> Both [violent and calm] passions pursue good, and avoid evil; and both of them are encreas'd or diminish'd by the encrease or diminution of the good or evil. But herein lies the difference betwixt them: The same good, when near, will cause a violent passion, which, when remote, produces only a calm one (T 2.3.4.1, SBN 419).

The most obvious interpretation of this last sentence is that a calm passion is one felt towards some object that is distant in time or space. Magri (2008, 198) understands Hume to be arguing that desires for distant objects are 'only minimally distorted by the accidental features' of their objects. This causes them to be calmer than desires for nearer objects, as well as a better guide to the 'greatest good' (Magri 2008, 198).

There are several problems with this interpretation, however. For one, there seems no good reason why evaluations of distant objects should be consistently preferable to evaluations of closer objects. There are surely other ways of correctly identifying an object's real features than by remaining at a literal distance from it. Moreover, Hume claims in this same passage that calm passions can be motivationally effective because they have been caused by 'repeated custom', and there is no obvious connection between custom and distance in space or time (T 2.3.4.1, SBN 419). We need an interpretation that allows for repeated

custom to produce the relevant kind of distance. A clue to this lies in Hume's definitions of calm and violent learned desires.

Hume's definition of a violent learned desire is as follows: 'Desire arises from good consider'd simply, and aversion is deriv'd from evil' (T 2.3.9.7, SBN 439). He defines a calm learned desire as 'the general appetite to good, and aversion to evil, consider'd merely as such' (T 2.3.3.8, SBN 417). The only substantive difference between the two definitions appears to be the addition of the word 'general' to the definition of a calm learned desire. At least, I take it that there is no important difference between considering something 'simply' and 'merely as such': Hume presumably means to indicate by these phrases that learned desires are directly and 'immediately' caused by perceptions of pleasure or pain, without requiring 'the conjunction of other qualities' (T 2.1.1.4, SBN 276).

These definitions suggest that violent desires are caused by particular ideas of pleasure or pain, whereas calmer desires are caused by more *general* ideas of pleasure or pain. Cohon (2010) argues for something similar. She focuses on Hume's discussion of 'strength of mind', in which he argues that the calm passions are at work whenever people 'counter-act a violent passion in prosecution of their interests and designs' (T 2.3.3.10, SBN 418). This suggests to Cohon (2010, 44), very reasonably, that the general appetite to good is 'the desire for one's own long-term well-being, or interest', which one will consider only in general terms. However, I will argue that the notion of generality plays a more central role within Hume's theory of calm passions than even Cohon suggests.

Hume does not describe calm pleasures or pains as 'general' in T 2.1.1.3 (SBN 276). However, I believe that the idea is implicit even here. To enjoy the particular auditory impressions as one listens to a song is to experience a violent pleasure. Thinking about the song later, where one is not presented with its particular features, one may feel a calmer pleasure. A yet calmer pleasure constitutes one's taste for that type of music generally. It is,

again, the level of generality of the perception which causes a passion that fundamentally determines whether the passion is a calm or a violent one.

*2.2.2. Calm passions as caused by general ideas*

McIntyre (2000, 83) notes several factors that Hume claims can increase the violence of a passion: 'proximity, especially in space or in the near future'; 'uncertainty, opposition'; 'novelty', and 'particularity'. However, McIntyre does not mention that, of these, only particularity – as opposed to 'generality' – is treated to a new section: 'Of the effects of custom' (T 2.3.5).

Hume begins this section by claiming that custom 'has two *original* effects upon the mind' (T 2.3.5.1, SBN 422). The first of these is to bestow a '*facility* in the performance of any action or the conception of any object' (T 2.3.5.1, SBN 422). To conceive of an object is to form an idea of it. Hume argues that we can form ideas of familiar things more easily than we can of unfamiliar things. To understand his argument, we must consider his theory of general ideas.

In Book 1, Hume argues that 'all general ideas are nothing but particular ones, annexed to a certain term, which gives them a more extensive signification, and makes them recall upon occasion other individuals, which are similar to them' (T 1.1.7.1, SBN 17). General or abstract ideas are the *same* mental objects as particular ideas, given a different function. We may, to use Hume's phrase, 'turn our view' to one idea in several ways, depending on the role it is playing within our thought at the time (T 1.1.7.18, SBN 25).

Hume seeks to explain the formation of abstract ideas by arguing that we often notice a 'resemblance' between certain ideas, so that we come to apply the same name to them and thereby categorise them together (T 1.1.7.7, SBN 20). Garrett (2002, 104) calls the groups in which ideas are 'annexed to a term' 'revival sets'. If I infer that you desire to help someone,

for example, then I will notice the similarity between my idea of your desire and my ideas of those desires called ‘generosity’. I will then place the idea of your desire in my revival set of generosity, and call this token desire ‘generosity’.

I can, of course, use the term ‘generosity’ to refer generally to desires of this type. When I do this, one of the particular ideas from the relevant revival set will come to mind and, somehow, represent all other ideas in that set, so that it will function as my abstract idea of generosity. Hume thus embraces the ‘paradox, that *some ideas are particular in their nature, but general in their representation*’ (T 1.1.7.10, SBN 22).

Hume’s theory allows for more subtle distinctions than just that between particular and abstract ideas. It suggests that we can turn our view to an idea in ways that make it neither fully abstract nor simply particular. To give a pertinent example for what is to follow, my idea of your generosity can be viewed not only as the idea of your particular motive or as the abstract idea of generosity, but also as the idea of your motive just *insofar as it is a token of generosity*. Indeed, we must view an idea in this way wherever we observe the resemblance between a particular idea of a motive and those ideas of motives called ‘generosity’, such that we classify it *as* a motive of generosity.

Following Hume, I will use the term ‘general notion’ to refer to any idea of an object, where the idea is viewed such that the object is considered merely as a token of some general type. Hume uses this phrase at only a few points: most notably, in discussions at T 2.3.6.2 (SBN 424) and T 2.3.6.4 (SBN 426), which we will soon consider, and during a discussion of moral judgement at T 3.3.3.2 (SBN 603), which we will consider in Chapter 6. Note that objects may be thought of as tokens of types that are more or less general: the general notion of a generous motive is in this sense a *less* general idea than the same idea when viewed as the general notion of a useful or agreeable motive.

To return to T 2.3.5.1 (SBN 422), I take Hume to be saying that, just as custom makes us more skilful at performing certain actions, so it makes us more skilful at forming general notions. Where an idea resembles others that we have frequently experienced, we find it easier to place it within a relevant revival set, and to name it, than we do where the idea is a relatively unusual one.

Hume then tells us about a further effect of custom: it creates in us a '*tendency or inclination*' to perform actions or conceptions with which we are familiar (T 2.3.5.1, SBN 422). Drawing on his previous point, this suggests that, whenever we encounter an object with very familiar features, so that it closely resembles objects that we have previously experienced, we have a strong tendency to place it in the relevant revival set, and to apply the relevant general term to it. When we come across anything familiar, we will habitually and quickly form a general notion of it as a token of its type. We may immediately categorise someone's motive as a 'generous' one, without requiring any conscious reflection to do so, but we would be unlikely to so easily categorise or name a token of a less common motive type, such as one of asceticism, for example.

In the next section, Hume claims that, '[w]herever our ideas of good or evil acquire a new vivacity, the passions become more violent; and keep pace with the imagination in all its variations' (T 2.3.6.1, SBN 424). This appears platitudinous: if I *believe* in a future pleasure, I will of course feel more violent joy than if I simply imagine it. This does not, however, seem to be the kind of distinction that Hume has in mind. His point is that 'the more general and universal any of our ideas are, the less influence they have upon the imagination' (T 2.3.6.2, SBN 425). If I form a 'particular and determinate idea' of some future pleasure, then I will feel a more violent joy than if I form an idea of that pleasure conceived only 'under the general notion of pleasure' (T 2.3.6.2, SBN 424). This seems to be because I have more to believe *about* the future pleasure when I have a more detailed, particular idea in mind than

when I have only a general notion. Ideas must be vivid to cause violent passions, and Hume appears to be suggesting that any particular and determinate idea is a highly complex one, comprising many vivid, simple ideas.

Hume's discussion of this is all very brief. Fortunately, he provides an example. He discusses a case from democratic Athens, in which the people were offered a vote on a proposed action, about which they were told only two things. They were told that it would benefit them all greatly in some way, and that it would be in some way unjust. The implication is that each Athenian felt two equally calm desires: one to gain some unknown benefit and the other to be just. They voted against the proposal which, Hume observes, may seem strange to many, because 'the advantage was immediate to the Athenians' (T 2.3.6.4, SBN 426). However, he claims that he can explain their actions:

> [Because the potential benefit] was known only under the general notion of advantage, without being conceiv'd by any particular idea, it must have had a less considerable influence on their imaginations, and have been a less violent temptation, than if they had been acquainted with all its circumstances (T 2.3.6.4, SBN 426).

The moral seems clear: general notions cause only calm passions. Only particular ideas have enough influence on the mind to cause violent passions. If the Athenians had formed an idea of some particular potential advantage, then they would have felt a violent desire for it, which would have required considerable 'strength of mind' to overcome (T 2.3.3.10, SBN 418). However, they had only the general notion of advantage, and this produced only a calm desire.

Recall that passions may be made violent by proximity, by uncertainty or opposition, or by novelty. All these are factors that are liable to make us focus on a particular object and its particular features. In contrast, the features that will make a passion calm are listed by McIntyre (2000, 83) as 'distance, especially in past time', 'security', and 'familiarity'. Hume seems to suggest that, wherever some object is long past, securely ours, or very familiar, we pay no close attention to it, so that we do not consider its particular features. We might think of something like a familiar painting that has long been in the hall, which we think of merely *as* that painting in the hall. Similarly, if we disbelieve that an object exists, or if we believe in an object in some distant land, then we will typically have fewer particular details in mind than when we consider nearby objects that we believe to exist. Indeed, many of our ideas of distant objects or events, and *most* of those of distant, future events, will be general notions: we will have nothing to form an idea *of* other than that which we generally associate with objects of the relevant type.

To summarise the key points of Hume's argument: an idea will typically produce a violent passion only where it is highly vivid, which typically or always requires it to be an idea of some particular and determinate object. Most violent passions are caused by beliefs about particular, nearby, pleasurable or painful objects. Wherever we turn our view to an idea such that it functions more generally than this, as with a general notion or an abstract idea, the idea will cause only a calm passion.

We saw in §2.1 that love is caused by a complex idea, comprising an idea of a person other than oneself and an idea of something closely related to her that causes a pleasure of some kind. As approbation is a calm form of love, it presumably has a similar cause, where one or more of the ideas within the complex idea is a general idea of some kind. In Chapter 3, I argue that Hume indicates where this idea is to be found, and how it causes approbation, in his discussions of 'delicate sympathy' (T 3.3.1.8, SBN 577).

# 3. A Calm and General Love: Hume's Theory of Approbation

This chapter draws on my arguments from Chapters 1 and 2 to demonstrate that Hume understands approbation as a calm form of love which, along with disapprobation, constitutes our taste in character traits.

In §3.1, I argue that Hume believes that all approbation is produced via a 'delicate' kind of sympathy, that responds to certain kinds of general notions, as discussed in §2.2.2. In §3.2, I argue that this allows Hume to distinguish the causes of approbation from those of love, so that he may consistently endorse theses (1), (2), and (3), as set out at the start of Chapter 1. We will see that Hume claims that approbation is always caused via sympathy with non-believed ideas of pleasure. In §3.3, I will argue that Hume can nevertheless allow that approbation may, in turn, produce motivationally efficacious desires, in accordance with his theory of motivation. Finally, §.3.4 addresses Hume's later moral *Enquiry.* I will argue that Hume's *Enquiry* account of approbation is consistent in all its fundamental details with his *Treatise* account.

## 3.1. The causes of approbation

Hume's arguments about the causes of approbation involve an account of what he calls 'delicate sympathy': 'Wherever an object has a tendency to produce pleasure in the possessor, or in other words, is the proper *cause* of pleasure, it is sure to please the spectator, by a delicate sympathy with the possessor' (T 3.3.1.8, SBN 576-7). This passage occurs in the course of his response to what Cohon (2010, 131) calls the '"virtue in rags" objection'. Hume sees this as a potential objection to his thesis that all moral sentiments are caused by sympathy. We approve of people who want to help others even where we know that they cannot help anyone, as where they are isolated in a 'dungeon or desart' (T 3.3.1.19, SBN

584). In such cases, there can be no beneficiaries with whom we may sympathise, so why do we approve?

Hume answers that sympathy produces approbation in such cases because it is readily influenced by custom: '*General rules* create a species of probability, which sometimes influences the judgment, and always the imagination' (T 3.3.1.20, SBN 585). This answer builds on arguments from throughout the *Treatise*.

### *3.1.1. General rules*

Hume's Book 1 account of the influence of general rules on the imagination is as follows:

> Our judgments concerning cause and effect are deriv'd from habit and experience; and when we have been accustom'd to see one object united to another, our imagination passes from the first to the second, by a natural transition, which precedes reflection, and which cannot be prevented by it (T 1.3.13.8, SBN 147).

As Hume notes, this is a description of a process of causal reasoning: 'all reasonings are nothing but the effects of custom; and custom has no influence, but by inlivening the imagination, and giving us a strong conception of any object' (T 1.3.13.11, SBN 149). However, general rules are those reasoning processes that occur rapidly and unreflectively.

Hume implies that we give the names 'sense and reason' to only those processes of general rules that ultimately produce reflectively endorsed beliefs (T 1.3.13.7, SBN 146-7). General rules often cause ideas to come to mind, and to become vivid, regardless of any reflective consideration of our situation. These ideas are often incompatible with what we call our 'beliefs', which are typically only those vivid ideas that we endorse once we *have*

reflected on our situation. Hume does not give a name to our unreflectively enlivened ideas, but I will call them ‘quasi-beliefs’.

It is because general rules produce quasi-beliefs that they are ‘the source of what we properly call Prejudice’ (T 1.3.13.7, SBN 146). Prejudice occurs wherever people form quasi-beliefs about the qualities of others, which continue to ‘influence their judgment, even contrary to present observation and experience’ (T 1.3.13.8, SBN 147):

> An *Irishman* cannot have wit, and a *Frenchman* cannot have solidity; for which reason, tho' the conversation of the former in any instance be visibly very agreeable, and of the latter very judicious, we have entertain'd such a prejudice against them, that they must be dunces or fops in spite of sense and reason (T 1.3.13.7, SBN 146-7).

Hume argues that quasi-beliefs are very often incompatible with our beliefs. In such cases, ‘our general rules are in a manner set in opposition to each other’ (T 1.3.13.12, SBN 149). For example, Hume describes how someone ‘hung out from a high tower in a cage of iron cannot forbear trembling, when he surveys the precipice below him, tho’ he knows himself to be perfectly secure from falling’ (T 1.3.13.10, SBN 148). The idea of falling is associated with the impression of the ground at a great distance below him, just as the idea of solidity is associated with the impression of iron. Both ideas come immediately to the person’s mind via custom, before he has any time to reflect. They are, however, incompatible as beliefs, because the solidity of the iron prevents falling.

Here, the person in the cage will very likely ‘correct’ for the idea of falling by ‘a reflection on the nature of [his] circumstances’ (T 1.3.13.10, SBN 148). In this way, he will reflectively endorse his vivid idea of his remaining safely within the cage. However, the

quasi-belief of falling will persist, and will retain some 'force and vivacity, which make it superior to the mere fictions of the fancy' (T 1.3.13.9, SBN 148). Part of its superiority, presumably, is that it can cause fear and 'trembling' (T 1.3.13.10, SBN 148).

This quasi-belief is, we must assume, a particular and detailed idea – that of falling from the cage in question – so that it will be a complex idea, comprising many simple ideas of what it would be like to fall. These simple ideas are all unbelieved, but each one is still vivid to some extent. Therefore, given Hume's understanding of the causes of violent passions, as discussed in §2.2, we can assume that the complex idea will produce a violent passion of fear. Moreover, the contrary belief cannot banish this complex idea from the mind: general rules precede our reflection and 'cannot be prevented by it' (T 1.3.13.8, SBN 147). In this way, Hume explains how we may feel fear in cases where we believe ourselves to be safe.

Although this person's quasi-belief and his belief in his safety within the cage are *both* vivid ideas within his imagination, he is likely to claim that he 'judges' that he is safe, but that he cannot help vividly 'imagining' falling. In such cases, the 'opposition of these two principles [of general rules] produces a contrariety in our thoughts, and causes us to ascribe the one inference to our judgment, and the other to our imagination' (T 1.3.13.11, SBN 149). There is, however, no fundamental difference in kind between a vivid, 'imagined' idea of falling and a 'reasoned' belief in one's safety.

In Book 2, Hume applies his theses of general rules and of quasi-beliefs to the operations of sympathy. He argues that they may cause us to feel passions that we do not *believe* to be felt by those with whom we sympathise. For example, where we encounter someone who is clearly not 'dejected by [her] misfortunes', we nevertheless associate her misfortunes with feelings of dejection, with which we then sympathise (T 2.2.7.5, SBN 370).

This is why we feel pained for those people who remain unperturbed in the face of misfortune, according to Hume.

*3.1.2. Hume's response to the 'virtue in rags' objection*

In Book 3, Hume explains cases of virtue in rags by arguing that we possess a 'delicate' kind of sympathy, which is so readily influenced by general rules that it ensures that we may approve of any character trait that is 'fitted to be beneficial to society', even where we do not believe that it will cause any happiness (T 3.3.1.20, SBN 585). We approve because we associate such traits with causing happiness, so that 'the imagination passes easily from the cause to the effect, without considering that there are still some circumstances wanting to render the cause a compleat one' (T 3.3.1.20, SBN 585).

If we infer that a person in a dungeon or desert has a benevolent motive, then, even if we know that she will never benefit anyone, we will habitually form a quasi-belief about the kind of happiness that benevolence typically causes. This idea will cause a sympathetic pleasure. We will then experience approbation, because the moral sentiments, like all 'sentiments of beauty', are 'mov'd by degrees of liveliness and strength, which are inferior to *belief*, and independent of the real existence of their objects' (T 3.3.1.20, SBN 585). This all appears to occur entirely unreflectively and habitually.

Unlike the man in the iron cage's quasi-belief about falling, the idea of happiness just discussed, with which we may sympathise, must be a general notion, as well as a quasi-belief. It is an idea caused merely by a general notion of benevolence, and it can only be conceived under the general notion of the kind of pleasure that benevolence typically causes. Any sympathetic pleasure that it causes will therefore be a calm passion.

In Book 3, Hume discusses several, non-moral cases where we may be influenced by delicate sympathy. For example, when silently reading some clumsy writing, we may

sympathise with the ‘uneasiness’ of someone reading it aloud, simply because we have often heard people struggle when reading such writing aloud (T 3.3.1.22, SBN 585-6). We presumably have no particular person in mind: the idea of uneasiness will be a general notion. Moreover, we will surely not so much as reflect on whether the idea of this uneasiness might represent anything real: the idea must be a quasi-belief. This quasi-believed general notion of uneasiness can only cause a *calm* sympathetic pain: a sentiment of taste, such that we judge the writing ‘harsh and disagreeable’ (T 3.3.1.22, SBN 586).

Hume then implies that the kind of sympathy just discussed is the *same* kind of ‘extensive’ sympathy ‘on which our sentiments of virtue depend’ (T 3.3.1.23, SBN 586). In Book 2, he has contrasted extensive sympathy with ‘limited’ sympathy (T 2.2.9.15, SBN 387). Limited sympathy is that which is caused by particular beliefs about a nearby person’s pains or pleasures, as felt in ‘the present moment’ (T 2.2.9.13, SBN 385). It will, therefore, typically produce violent sympathetic pains or pleasures.

In Book 2, Hume’s main discussion of extensive sympathy concerns cases where we form beliefs about a person’s life, beyond that which is immediately present to us (T 2.2.9.14, SBN 386). If I see a child in poverty, then I may feel a violent pain via limited sympathy. I may also form beliefs about her unseen family, or about her future hardships, and I will then feel further sympathetic pains, via extensive sympathy.

As we have seen, delicate sympathy is a different kind of extensive sympathy from that discussed at T 2.2.9.14. Delicate sympathy is that kind which occurs where we sympathise with quasi-believed general notions of pain or pleasure. Any sympathetic pleasure or pain so produced can only be a calm passion. (For simplicity, I will henceforth generally refer to quasi-believed general notions merely as ‘quasi-beliefs’. Although some quasi-beliefs may be particular ideas, and so cause violent passions, we need not consider these further.)

In Book 3, Hume mainly discusses delicate sympathy in relation to aesthetic sentiments, and in his response to the virtue in rags objection. It is therefore tempting to understand him to argue that, although approbation is typically produced via limited sympathy, it is in some atypical cases produced via delicate sympathy. However, we saw that he implies that our 'sentiments of virtue' *depend* on delicate sympathy (T 3.3.1.23, SBN 586). Delicate sympathy always causes calm passions, whereas limited sympathy typically causes violent passions, as do many cases of non-delicate, extensive sympathy. Moreover, Hume *defines* 'virtuous' character traits as those that cause 'pleasure by the mere survey' by being 'naturally fitted' to be useful or agreeable (T 3.3.1.30, SBN 591). In the discussion leading up to this definition, he has identified only one process – that by which we experience a delicate sympathy with quasi-beliefs about pleasure – which causes us to automatically feel a calm pleasure towards just those traits that are 'fitted to be beneficial to society' (T 3.3.1.20, SBN 585). This suggests that he intends us to understand this as the kind of process by which approbation is always caused.

Clearly, we do not think that a benevolent person in a dungeon has a less virtuous motive than a similarly benevolent but free person, just because she cannot act on her benevolent desires. Moreover, we do not need to reflect on or compare these two motives in any detail if we are to evaluate these two characters equally. We simply approve of anyone whom we take to be benevolent. Hume relies on his accounts of general rules, quasi-beliefs, and delicate sympathy to explain this: whenever we see someone with a benevolent motive, we habitually form an idea of the kind of happiness that benevolence typically causes, sympathise with it, and then experience a strong, but calm, approbation. This all suggests, I think rightly, that Hume intends to argue that *all* instances of approbation are caused purely via a process whereby token character traits habitually and unreflectively cause quasi-beliefs about pleasure, with which we sympathise via delicate sympathy.

Hume thus argues for a distinction, but also a strong parallel, between reflective causal reasoning and the kind of unreflective process that produces a moral sentiment. Moral judgements are caused via delicate sympathy with quasi-beliefs, and they rely on no reflectively endorsed beliefs about the effects of the traits that cause them. However, the process by which a quasi-belief is formed is itself a process of the customary association of ideas, and of the transference of vivacity, as is any process of reasoning. Although moral sentiments are not formed by processes of reflectively endorsed reasoning, they *are* formed by processes that are very similar, and just as reliant on experience, as those that are reflectively endorsed. The judgements that these processes produce are not ideas that feel believed, but rather passions that feel like believed ideas.

If all moral judgements are produced by delicate sympathy, as suggested above, then Hume must endorse the following thesis:

GENERALITY: All ideas of typically beneficial or pleasing character traits habitually cause approbation, all ideas of typically harmful or displeasing character traits habitually cause disapprobation, and the strength of any moral sentiment is dependent only on the degree of happiness or unhappiness with which the type of trait is generally associated.

According to Generality, the moral sentiments are uniform:

UNIFORMITY: Any kind of evaluative or psychological response is *uniform* if and only if it is such as to respond in the same way towards all token character traits of any one type,

> regardless of how the responder is related to the person whose trait it is, or is affected by the particular effects of the token trait.

These are not, of course, Hume's terms. I will call a kind of evaluative or psychological response 'variable' where it is such as to respond differently towards different tokens of the same general type of character trait, depending on one's relation to the person whose trait it is, or on the particular effects that the token trait has on oneself. According to Generality, approbation is uniform rather than variable: *any* motive of benevolence, for example, will produce a strong but calm, sentiment of approbation, regardless of our beliefs about its context or likely effects, via a purely habitual and automatic process.

Unfortunately, rather than arguing for this directly, Hume argues more generally that all sentiments of taste, not just sentiments of moral taste, are caused by general rules and quasi-beliefs. His core thesis is that the '*seeming tendencies* of objects affect the mind: And the emotions they excite are of a like species with those, which proceed from the *real consequences* of objects, but their feeling is different' (T 3.3.1.23, SBN 586). General rules very often produce quasi-beliefs about the typical consequences of token objects, which then cause calm passions. Any such passion will be of the same kind that we would also feel violently if we were to believe in such consequences. For example, wherever we see a building that 'seems clumsy and tottering to the eye', we experience a calm 'fear', which causes (or, perhaps, *is*) a 'sentiment of disapprobation', such that we find the building 'ugly' (T 3.3.1.23, SBN 586).

No delicate sympathy is involved in causing the kind of sentiment just discussed. Although Hume *does* discuss delicate sympathy in this paragraph, he does so only to argue that we may feel a calm 'pain and disapprobation' even where we feel no violent pains via

limited sympathy (T 3.3.1.23, SBN 586). He does not piece together his arguments concerning approbation in any one section. However, we may now do so on his behalf.

### 3.2 Reconciling Hume's claims about approbation

I began Chapter 1 by summarising three core theses within Hume's theory of approbation, as follows: (1) Approbation is a calm form of the 'indirect' passion of love; (2) Approbation is a sentiment of taste, in virtue of which we come to value useful and agreeable kinds of character traits; (3) Approbation is in all cases caused, via sympathy, by an idea of a character trait with a 'tendency to the good of mankind'. We are now in a position to reconcile these claims.

We have seen that love is that violent pleasure caused by a complex idea of some person other than oneself, and of something pleasing that is closely related to that person by causation or contiguity. Approbation is a calm pleasure that is similarly caused by a complex idea, comprising an idea of a person and an idea of some character trait of hers that causes one to feel a calm pleasure, via delicate sympathy with a quasi-belief about the type of pleasure that one associates with such traits.

We approve of any 'character' that is 'naturally fitted' to cause happiness (T 3.3.1.30, SBN 591). This is best understood as a token character *trait* of some generally pleasing type. However, Hume clearly assumes that we regard any such trait as a *person's* trait. This is understandable enough, since any idea of a character trait will be strongly associated with that of a person. Any complex idea that causes approbation therefore includes at least an idea of a character trait, an idea of some person whose trait it is, and – because the relevant trait is one that we take to be 'naturally fitted' to cause happiness – an idea of the kind of happiness that such character traits generally cause.

We saw in §2.1.1 that Hume often talks of being proud of some possession, such as a beautiful house, where he means that one is proud of oneself as the owner of that possession. We saw in §1.3 that Hume claims that the object of approbation is always a motive, or other character trait, but that he nevertheless frequently discusses our approval of people. In both cases, Hume appears to simply assume the idea of a person. Just as pride is always caused by a complex idea of oneself as the owner of a pleasing possession, it appears that approbation is always caused by the complex idea of a person with a motive, or other character trait, of some typically pleasing kind.

For any such complex idea to cause approbation, the idea of the *motive* must be a general notion, or a general idea of some other kind, because only general ideas can cause calm passions via delicate sympathy. The idea of the *person* may presumably be either general or particular. If we form the abstract idea of a benevolent person, then general rules and delicate sympathy will, presumably, ensure that we approve of this general kind of person. This is, presumably, what happens whenever we sincerely say, 'benevolent people are virtuous'.

If we form an idea of a *particular* benevolent person, then we can turn our view to the idea of her motive in at least two ways. We can view it as the idea of the particular motive of the particular person, such that we may feel pleasure (whether via limited sympathy or more directly) from our impressions of, or beliefs about, its particular effects. Viewed in this way, our idea of a person with a benevolent motive may or may not cause the violent passion of love, depending on the context. However, we are also very likely to view the idea as a general notion, as we will habitually form a conception of the motive as a 'benevolent' one. We will then habitually form a quasi-belief about the kind of happiness that benevolence typically causes, sympathise with it, and experience approbation.

This explains the double relation of ideas and impressions that causes approbation. The relation of ideas occurs where a general notion (or other general idea) of a person's motive is associated with an idea of the happiness that such motives typically cause. These associated ideas combine to form a complex idea of a person with a motive that causes happiness. The relatively simple (quasi-believed) idea of happiness produces a calm pleasure, via delicate sympathy. The more complex idea then causes a calm pleasure of approbation, related to the sympathetic pleasure by resemblance. Approbation is always caused in this way. This gives us (1): Approbation is a calm form of the 'indirect' passion of love. It also gives us (3): Approbation is in all cases caused, via sympathy, by an idea of a character trait with a 'tendency to the good of mankind'.

Just as a calm fear will occur wherever we see a seemingly tottering building, regardless of our beliefs, any token motive of a generally pleasing type will cause approbation, regardless of our beliefs: 'The imagination adheres to the *general* views of things, and distinguishes betwixt the feelings they produce, and those which arise from our particular and momentary situation' (T 3.3.1.23, SBN 586). By this, Hume means that we can distinguish our calm sentiments of taste, caused by quasi-believed general notions, from our violent passions of the same kind, caused by particular beliefs, even where these occur together.[1] I may disapprove of someone for her cruelty even as I love her for her courage, and I will not confuse the calm sentiment with the violent passion. Indeed, I may simultaneously *love* her for her courage, where I view the idea of her courage as a particular idea, even as I approve of her for her courage, where the same idea is viewed as a general notion.

[1] Any quasi-belief of the building falling that caused a *violent* fear would have to be a particular idea, and so also distinct from the quasi-believed general notion that causes a calm fear.

Approbation may therefore be clearly distinguished from love. Along with disapprobation, it constitutes our taste in characters, character traits and, derivatively, in actions. This gives us (2): Approbation is a sentiment of taste, in virtue of which we come to value useful and agreeable kinds of character traits.

Hume believes that we approve more of virtues that we strongly associate with causing great happiness, like justice, than we do of others (T 3.3.1.9, SBN 577). He also suggests that 'natural abilities' produce an '*inferior*' 'sentiment of approbation' than that caused by what we typically think of as 'virtues' (T 3.3.4.2, SBN 607). He generally allows that approbation feels different when caused by different kinds of virtue, and he explains this by reference to the kinds of *non-moral* pleasure that we associate with each kind of trait. Some kinds of traits, like Caesar's, are pleasurable in such a way that they produce 'love', as the passion is usually understood (T 3.3.4.2, SBN 608). Other kinds of traits, like Cato's, are pleasurable in a 'severe and serious' way, and they are such as to produce a form of love which Hume calls 'esteem', and which seems to be something like respect (T 3.3.4.2n88, SBN 608).[2] Hume claims that we approve of both '*Cæsar* and *Cato*', but that the approbation feels different in each case (T 3.3.4.2, SBN 607; see also M App. 4.6, SBN 316-7). Just as our taste in music consists of calm forms of the kinds of passions that we feel towards particular pieces of music, so our taste in character traits consists of calm forms of the kinds of passions of love and hatred that we feel towards particular traits.

However, throughout the *Treatise*, Hume generally assumes that we are all affected and influenced in the same ways by our ideas of the same kinds of characters, traits and actions. He does not seem to allow that anyone might not approve of all generally useful or

[2] This is how Hume uses the term 'esteem' throughout Book 2. He frequently refers to 'love and esteem' in his discussions of love, as at T 2.2.1.4 (SBN 330). Confusingly, however, in Book 3, Hume often uses 'esteem' to mean 'approbation', instead of 'respect', as at T 3.3.1.14 (SBN 580-1).

agreeable traits, and he is largely oblivious to moral disagreement. He thinks we need not worry about whether we can talk of 'a *right* or a *wrong* taste in morals', because 'there is such an uniformity in the *general* sentiments of mankind, as to render such questions of but small importance' (T 3.2.8.8n. 80, SBN 547).

I believe that the theory of approbation, as described so far, constitutes most of what Hume takes to be his substantive contribution to 'our reasonings concerning *morals*' (T 3.1.1.1, SBN 455). He has, he thinks, explained what moral judgements are: calm forms of the passions of love and hatred, which constitute our sentiments of taste in characters. He has, he thinks, explained how they are caused: via the operations of general rules and of our delicate sympathy with the quasi-beliefs so produced. And he has, he thinks, explained what it is about character traits that make them such as to cause approbation or disapprobation: they are all and only those traits that we generally associate with causing happiness or unhappiness to their possessors or to those people around them. These explanations all rely on, and cohere with, various theses from Books 1 and 2 of his *Treatise*: they 'corroborate whatever has been said concerning the *understanding* and the *passions*' (T 3.1.1.1, SBN 455).

However, most of Hume's readers are at least as interested in at least three further aspects of his treatment of 'morals': his distinction between natural and artificial virtues; his arguments that moral judgements are not produced by reason alone; and his discussion of the common or general point of view which, he claims, we adopt when we moralise. I will discuss the first of these in Chapter 4, the second in Chapter 5, and the third in Chapter 6. First, however, there are two further points to address. In §3.3, I will discuss Hume's theory of motivation, to demonstrate that it is compatible with the foregoing. Then, in §3.4, I will consider Hume's account of approbation in his moral *Enquiry*.

### 3.3. Hume's theory of motivation

Of Hume's many influential theories, his theory of motivation has probably exerted the greatest influence on contemporary metaethics. In the late 20th century, Smith (1994, 92) described the 'Humean theory of motivation' as a contemporary 'dogma in philosophical psychology'. This theory, in its strongest sense, is committed to the claim that all 'motivation has its source in the presence of a desire and means-end belief', where these two mental states are modally distinct from one another (Smith 1994, 92). In this section, I will argue that, so construed, it is not Hume's view.[3] This is because, I will argue, Hume allows for quasi-beliefs to play the role of beliefs in many cases of motivation.

For Hume, any motive is an 'impulse of passion': a desire (T 2.3.3.4, SBN 415). As we saw in §2.2, most desires are caused by the 'perception of pain and pleasure', which produces aversions for painful objects and desires for pleasurable ones (T 1.3.10.2, SBN 118). This form of motivation may occur when we experience a feeling of pain or pleasure, as when we touch a painfully hot surface. Alternatively, it may occur when 'we have the prospect of pain or pleasure from any object' (T 2.3.3.3, SBN 414). For example, if I believe that a surface would be painful to touch then, according to Hume, this belief will motivate me only in conjunction with a distinct aversion to that pain.

According to the strongest version of the Humean theory of motivation, as applied to his own philosophy, Hume denies that beliefs either do or could cause desires or motivating passions. For example, Radcliffe (1999, 101) understands Hume to deny this, because a 'rationalist might well agree that actions are not caused by beliefs alone'. Given Hume's clear opposition to moral rationalism, she thinks that Hume must intend to assert something that rationalists could not agree with. She therefore argues that Hume's theory of motivation

[3] Here, I contribute to a longstanding debate. See, for example, Millgram (1995) and Persson (1997).

entails ‘the incapacity of reason to generate the motivating passions in the first place’ (Radcliffe 1999, 101). A belief that, for example, the ice cream before me would be pleasurable to eat *cannot* be the proximate cause of my desire to eat the ice cream, according to Radcliffe.

This interpretation is a contested one. For example, Cohon argues that Hume merely intended to deny ‘that a causal belief that is not about *pleasure or pain* can produce a desire or aversion. Hedonic beliefs presumably can’ (Cohon 2010, 50, Cohon’s emphasis). Consider his claim that ‘when we have the prospect of pain or pleasure from any object, we feel a consequent emotion of aversion or propensity, and are carry'd to avoid or embrace what will give us this uneasiness or satisfaction’ (T 2.3.3.3, SBN 414). This strongly supports Cohon’s argument, for here Hume clearly sees an emotion as being caused by a prospect of pain or pleasure.

Our question now is what, precisely, Hume means by a ‘prospect’ of pain or pleasure. The phrase occurs during his arguments to show that ‘reason alone can never be a motive to any action of the will’ (T 2.3.3.1, SBN 413). Certainly, therefore, we should assume that many such prospects are, according to Hume, reasoned *beliefs* about pain or pleasure. If Hume believes that *all* such prospects are beliefs, then he endorses a version of the Humean theory of motivation. My motivation to eat the ice cream is due to my belief that the ice cream will taste nice and to my desire to eat the ice cream, where the desire’s proximate cause is the belief. Certainly, the belief and the desire are modally distinct, according to Hume (T 1.4.6.16, SBN 259).

Few readers of Hume believe that he allows for any cases of motivation to be due only to desires and non-believed ideas of pleasures and pain. Cohon (2010, 42n. 11) takes him to claim that unbelieved ideas can cause desires, but that belief is ‘needed… to enable the desire to engage the will (cause action)’. The motivational inefficacy of unbelieved ideas

is often assumed, rather than explicitly argued for. For example, Persson (1997, 195) claims that reasoning influences action by having an 'impact on belief and passion'. Similarly, Darwall (1993, 423) claims that ideas of pain or pleasure may influence action where they are 'sufficient in force and vivacity to constitute beliefs'.

I think this assumption is sufficiently widespread to call it the 'standard interpretation' of Hume's thesis of the relationship between vivacity, ideas and motivation. If the standard interpretation is correct, then my interpretation of Hume must be false, because I read him as arguing that no beliefs about pleasure are involved in the production of approbation. According to the standard interpretation, he cannot allow that any approbation strong enough to cause a motivationally effective desire could be caused by anything less than a *believed* idea of happiness, as caused by some relevant motive.

Fortunately, we have several good reasons to think the assumption false, and so to disagree with the standard interpretation. For one, Hume claims that passions may be caused by impressions or ideas (T 2.1.1.1, SBN 275). He does not specify that they are only caused by impressions and *believed* ideas.

Hume appears to allow in several places that quasi-beliefs about other people's pleasures and pains can produce motivationally influential pleasures or pains in us, via habit and sympathy. True, belief is '*almost* absolutely requisite to the exciting our passions' (T 1.3.10.4, SBN 120, my italics). However, a close examination of Hume's theory of causal reasoning demonstrates that he allows for several passions to be 'excited' by 'prospects' of pain and pleasure that are *not* beliefs.

Consider again Hume's account of delicate sympathy and taste, discussed in §3.1, which allows that our sympathy with a quasi-belief about pain may cause the judgement that a book is poorly written. Similarly, Hume argues that a house may be deemed beautiful because it 'is contriv'd with great judgment for all the commodities of life [and] pleases us

upon that account; tho' perhaps we are sensible, that no-one will ever dwell in it' (T 3.3.1.20, SBN 584). Clearly, we may be motivated to discard a book because it is displeasing, or to take a picture of a house because it is beautiful. Therefore, Hume must allow that these aesthetic judgements may be strong enough to produce motivationally effective desires.

Consider too Hume's discussions, in Book 2, of the effects of general rules on sympathy, also discussed in §3.1. He argues that we may sympathise not only with the non-existent pains of someone 'who is not dejected by misfortunes', but also of those who 'behave themselves foolishly before us' without being 'in the least conscious of their folly' (T 2.2.7.5, SBN 370-1). Surely our sympathies here may motivate us, to assist the former or, perhaps, to have a quiet word with the latter. Similarly, it is only sympathy with obviously non-existent pains that can make us feel compassion with those who are murdered when asleep or with children in perilous situations that they are too young to comprehend (T 2.2.7.6, SBN 371). Clearly, we may be motivated by such compassionate feelings.

Finally, we saw that the man in the iron cage felt fear as a consequence of his quasi-belief about falling (T 1.3.13.10, SBN 148). We surely ought to allow that *this* passion is a motivating one, for a great many people in this person's position would be moved by their fear to leave. Hume does not say as much, but he certainly implies that the man in the cage may be so motivated:

> His imagination runs away with its object, and excites a passion [of fear] proportion'd to it. That passion returns back upon the imagination and inlivens the idea; which lively idea has a new influence on the passion, and in its turn augments its force and violence; and both his fancy and affections, thus mutually supporting each other, cause the whole to have a very great influence upon him (T 1.3.13.10, SBN 148-9).

It would be both surprising and wholly implausible if Hume were to deny that the 'very great influence' of fear could not, in this case, be motivationally effective, just because the man in question does not believe that he will fall.

Certainly, none of the textual evidence here supports the assumption that Hume requires belief for all cases of motivation. Although he does not say so explicitly, there appear to be many cases of motivation that he aims to explain by reference to quasi-beliefs causing motivating passions. In Chapter 4, I will argue that, given my interpretation of Hume's theories of moral judgements and motivation, he therefore has a simpler and more plausible explanation for our consistent approval of justice than can otherwise be attributed to him.

For this to be a plausible interpretation of Hume, however, it is not sufficient to show that he allows for quasi-beliefs to take the place of beliefs in moral motivation: I must also show that the moral sentiments can be indirectly motivating, by being such as to cause desires. Of course, it is precisely their 'influence on human passions and actions' which demonstrates to Hume that they are passions, rather than the conclusions of reason (T 3.1.1.5, SBN 457). Clearly, Hume thinks we *are* motivated by our moral sentiments. The important question now is: how do we reconcile this claim with his belief that they are calm forms of love and hatred?

This may appear to be a very straightforward question to answer, because we have seen, in §2.1.1, that Hume tells us that love and hatred consistently cause desires for the happiness or unhappiness of their objects (T 2.2.6.3, SBN 367). This certainly explains why we are motivated to reward the virtuous and to punish the vicious, according to Hume. However, the explanation is a slightly more complicated one than we might expect.

Hume claims that 'the most obvious causes' of love and hatred are virtue and vice: wherever we approve or disapprove of someone, we then feel love or hatred towards her (T

2.1.7.2, SBN 295). Our desires to reward or punish such people are at least primarily caused only by these *non-moral* forms of love or hatred:

> As to the good or ill desert of virtue or vice, 'tis an evident consequence of the sentiments of pleasure or uneasiness. These [moral] sentiments produce love or hatred; and love or hatred, by the original constitution of human passion, is attended with benevolence or anger; that is, with a desire of making happy the person we love, and miserable the person we hate (T 3.3.1.31, SBN 591).

Here, then, is Hume's explanation of our motives to reward the virtuous and to punish the vicious. However, this only appears to explain why we feel these desires where we take *other* people to be virtuous or vicious. We saw, in Chapter 1, that he denies that there is truly any self-love or self -hatred. This suggests that self-approbation and self-disapprobation are forms of pride and humility. However, Hume cannot claim that humility makes us desire to punish ourselves for our viciousness, in the way that hatred makes us want to punish vicious others. This is because he denies that humility causes any such desires:

> The passions of love and hatred are always followed by, or rather conjoin'd with benevolence and anger. 'Tis this conjunction, which chiefly distinguishes these affections from pride and humility. For pride and humility are pure emotions in the soul, unattended with any desire, and not immediately exciting us to action (T 2.2.6.3, SBN 367).

It is not entirely clear how, or if, Hume allows that we would be motivated to seek rewards or punishments for ourselves, on account of our own moral characters as we perceive them. I suspect that, in fact, Hume sees *all* instances of approbation and disapprobation as calm forms of love and hatred, even when directed towards oneself. He argues that we may come to 'hate' ourselves when we lack certain virtues, and so come to feel a desire to act as if we possessed these virtues (T 3.2.1.8, SBN 479). It appears, therefore, that any disapprobation felt towards oneself is a calm form of hatred, rather than of humility. Self-*love* is no real passion, according to Hume, but it appears that self-*approbation* is.

Of course, motives to punish and reward people are not the only morally relevant motives. Hume must also explain why we are frequently motivated to act in accordance with our moral sentiments: why we often act because it is virtuous to do so, or because it would be vicious not to do so. He does not give a detailed explanation for this, but I think he assumes that all double-relation passions may cause desires, simply because they are pleasures and pains. Admittedly, as a 'pure emotion', pride *alone* cannot motivate: it is not itself a desire and it does not consistently produce any one type of desire, in the way that love consistently produces benevolence. Nevertheless, it seems entirely consistent and plausible to allow that the thought of owning something that would increase my pride *would* be produce a desire to own that object. As a pleasure, pride may cause desires, even if it does not consistently cause any one kind of desire.

Similarly, approbation and disapprobation are forms of pleasure and pain (T 3.1.2.3, SBN 471). They are, therefore, such as to cause desires and aversions, as when we desire to perform an action because we approve of the idea of it. We may take it, therefore, that wherever we experience approbation, this passion will typically produce a desire to perform any relevant action, or to promote the trait or action in question.

Hume's theory of motivation is, therefore, compatible with my interpretation of Hume's theory of moral sentiment. I will now turn from Hume's *Treatise* to his moral *Enquiry*, to examine the details of his account of approbation in that work.

### 3.4. Hume's theory of moral judgements in his moral *Enquiry*

In this section, I argue that Hume's theory of approbation in the *Enquiry* is largely consistent with that in the *Treatise*. However, this is not always immediately clear, and Hume does not appear greatly interested in vindicating the precise details of his theory in this later work. His main aim is to apply the 'experimental method' to morality, by gathering evidence from history, literature, and our common experiences surrounding moral practice, to demonstrate that we typically approve of generally useful or agreeable character traits and disapprove of generally harmful or disagreeable ones (M 1.10, SBN 174).

Throughout most of the *Enquiry*, Hume is officially neutral on the question of whether moral judgements are beliefs or sentiments. Unfortunately, one consequence of this is that he typically refers to violent passion by terms like 'real feeling or sentiment' and to calm passions by terms like 'general judgments' (M 5.41n. 24.1, SBN 228).[4] Nevertheless, his moral sentimentalism is thinly disguised: he cannot resist describing moral judgements as 'calm passions and propensities' (M 6.15, SBN 239). He ultimately defines virtue as '*whatever mental action or quality gives to a spectator the pleasing sentiment of approbation*' (M App. 1.10, SBN 289).

We will see, in Chapter 6, that Hume greatly improves on his *Treatise* account of moral language in this work. There is only one further, important difference: Hume allows

[4] In Chapter 6, we will see that he also uses common language in this way, and with similarly confusing consequences, in his *Treatise* account of the common point of view.

for greater variation in our moral sentiments in the *Enquiry*. He allows that different moral norms hold in different cultures, although he thinks that this can be entirely explained by the fact that different characters and actions may be useful or agreeable in different cultural contexts (M D.37, SBN 336). As a more fundamental change, he now allows that at least some individuals within any one culture might contemplate some tokens of generally pleasing traits *without* experiencing strong enough moral sentiments to motivate them. A 'sensible knave' might recognise that it is always just to pay back a loan, but then happily decide not to return some money if she believes that little harm will be caused by this (M 9.22, SBN 282). In the *Enquiry*, therefore, it appears that Hume treats Generality as only typically true, rather than as universally true. This aside, I will argue Hume commits to Generality and to his theory of delicate sympathy in this work.

Capaldi (1989, 240) argues that Hume appears to reject the *Treatise* account of sympathy, when he claims that a '*real* sentiment or passion' could not plausibly 'arise from a known *imaginary* interest' (M 5.13, SBN 217). However, as Vitz (2004, 270) argues in response, in the relevant passage Hume is 'merely claiming that the pleasure that people get from utility is not derived solely from an imaginary self-interest'. Hume generally understands our violent, 'real' passions to be broadly self-interested ones, unlike our moral sentiments. Here, his point is that we do not call distant characters 'virtuous' just because we imagine how they would benefit us or our loved ones if they were nearer.

As in the *Treatise*, Hume claims that we feel sympathetic pleasures at the pleasures of others and sympathetic pains at their pains (e.g. M 5.23, SBN 221). Moral sentiments arise from the 'principles of humanity and sympathy' (M 5.45, SBN 231). Moreover, as Debes (2007a, 38-39) argues, the *Enquiry* account of justice remains consistent with the *Treatise* account: Hume argues, as we will see in Chapter 4, that our general approval of justice develops from a combination of artifice, self-interest, and sympathy. I therefore agree with

Debes (2007a; 2007b), along with others such as Abramson (2001), Penelhum (1975, 147-148), and Vitz (2004), that we should understand Hume to tacitly refer to sympathy when he uses terms like 'humanity' and 'benevolence' in the *Enquiry*.

Hume is often imprecise in his use of the term 'humanity', but – particularly in later parts of the *Enquiry* – he frequently uses it as a shorthand for the 'principle of humanity' (M 9.6, SBN 272). I will follow this usage, and I will argue that this principle just is that of delicate sympathy, as it responds to quasi-beliefs about pleasure or pain. It is humanity which gives us a 'general approbation of what is useful to society, and blame of what is dangerous or pernicious' (M 5.39, SBN 226). This is precisely the role that delicate sympathy plays in the *Treatise.* Hume sometimes also calls humanity 'general benevolence' (M App. 2.5n. 60.1, SBN 300). Again, he is not always consistent in his use of this term, but 'benevolence' in the *Enquiry* is, at least sometimes, a synonym for sympathy (rather than for the desire to help those around us, as in the *Treatise*).

Unlike Taylor (2013), I understand humanity to respond to agreeable traits, as well as to useful ones. Indeed, the *Enquiry's* most explicit example of the influence of humanity on moral sentiment has it operating on traits that are immediately agreeable to others (M 8.15, SBN 267).[5]

As in the *Treatise*, 'delicate' sympathy responds to unbelieved, habitually produced ideas of pleasure or pain (M 5.37, SBN 224). Indeed, the *Enquiry* provides further evidence that the kind of extensive sympathy that produces approbation just is delicate sympathy. Hume explicitly says that it is 'delicate' sympathy that causes us to find any 'unharmonious composition' of writing 'harsh and disagreeable' (M 5.37, SBN 224). In the *Treatise*, he only uses the more general label 'extensive sympathy' to describe this kind of response, although

[5] Reed (2016, 1151-1152) makes a similar observation.

there he explicitly says that it is the kind of sympathy that causes approbation (T 3.3.1.22, SBN 585-6; T 3.3.1.23, SBN 586).

As in the *Treatise*, this example, of sympathising with an imagined reader, helps Hume to explain and defend his account of the causes of moral sentiments. It is analogous to his claim that, wherever we associate types of character trait with causing pain or pleasure to others, *any* token trait of any such type brings a quasi-belief about pain or pleasure to mind, with which we sympathise. These sympathetic pains and pleasures then produce moral sentiments. Hume gives a clear example of this process regarding immediately agreeable traits:

> We approve of another, because of his wit, politeness, modesty, decency, or any agreeable quality which he possesses; although he be not of our acquaintance, nor has ever given us any entertainment, by means of these accomplishments. The idea, which we form of their effect on his acquaintance, has an agreeable influence on our imagination, and gives us the sentiment of approbation (M 8.15, SBN 267).

Hume tells us that this 'principle enters into all the judgments, which we form concerning manners and characters' (M 8.15, SBN 267). This is strong evidence that he still endorses Generality.

In the *Enquiry*, Hume stresses that our *verbal* moral evaluations are uniform: he thinks that 'in every discourse and conversation' people can be seen to praise the same types of traits, such as 'honour', 'wit' and 'gallantry', even where they are personally unaffected by them (M 9.2, SBN 269). He argues that the observable uniformity of our verbal moral evaluations is evidence that approbation is uniform: 'The notion of morals, implies some

sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it' (M 9.5, SBN 272). No matter how near or far we are to a character, even one 'the most remote' from us, Hume thinks that we will mostly agree in our moral judgement of it. The sentiment responsible for this uniformity in judgement is 'the sentiment of humanity': that kind of sentiment that arises from the principle of humanity, and so a moral sentiment (M 9.5, SBN 272).

The *Enquiry* implicitly endorses the *Treatise* view that approbation is a calm form of love (T 3.3.5.1, SBN 614). Hume argues, mainly via an unpleasant discussion of an 'untaught savage', that our non-moral forms of love and hatred are caused by non-delicate sympathy, such that we sympathise with the pleasures or pains of individuals with whom we have some connection (M 9.8n. 57.1, SBN 274-5). Hume claims that the untaught person 'regulates chiefly his love and hatred by the ideas of private utility and injury, and has but faint conceptions of a general rule or system of behaviour' (M 9.8n. 57.1, SBN 274). Any such person who sees his friend struck down in battle will feel only hatred towards his friend's assailant, where this hatred is caused by 'particular' benevolence: a sympathetic reaction to the pain of someone with whom he has a friendly relationship (M App. 2.5n. 60.1, SBN 298). Humanity is absent here: this person lacks any delicate sympathy or 'general' benevolence, so that he cannot experience moral sentiments (M App. 2.5n. 60.1, SBN 298).

The non-savage, in contrast, is 'accustomed to society, and to more enlarged reflections', so that she is aware that fighting for one's country tends to produce happiness for one's compatriots, whatever one's nationality (M 9.8n. 57.1, SBN 274). She therefore experiences 'a general sympathy', or humanity, with the beneficiaries of such happiness, so that she feels approbation towards her opponents (M App 2.5n. 60.1, SBN 298). She may

also, simultaneously, experience 'ruder and narrower passions', by violently hating those who harm her friends, even as she approves of their valour (M 9.8n. 57.1, SBN 275).

The moral and non-moral kinds of love and hatred are therefore distinguished by the kinds of sympathy that produce them, as in the *Treatise*. Where we contemplate a motive's particular effects, we experience particular benevolence and a non-moral sentiment, such as love. Where we contemplate a motive just as a token of its general type, we experience general benevolence and a moral sentiment, such as 'general approbation' (M 5.45, SBN 231). These processes may occur simultaneously as we contemplate any one motive.

However, Hume uses an unfortunate phrase, which appears to suggest that they cannot occur together. He describes 'cool approbation' being '*converted* into the warmest sentiments of friendship and regard' as we get to know someone better (M 5.43, SBN 230, my emphasis). Presumably, he means that our approbation continues in such cases, but now alongside the warmer sentiments. No plausible theory of moral judgement could allow that we cease to approve of someone just because we come to love them. The unfortunate phrase is presumably only intended to demonstrate that violent love and calm approbation have very similar, sympathetic, causes.

Hume never says that we only approve or disapprove of motives that we *believe* will benefit or harm others. Instead, he stresses that we approve or disapprove of motives with a *tendency* to benefit or harm others.[6] This phrasing is clearly compatible with, and suggestive of, Generality. Similarly, while Hume claims that we approve of only useful or agreeable traits, he describes utility itself as 'only a tendency to a certain end' (M App. 1.3, SBN 286).

[6] M 2.22, SBN 181; M 3.8, SBN 186; M 3.24, SBN 193; M 3.40, SBN 201; M 5.4, SBN 214; M 5.16, SBN 218; M 5.39, SBN 225; M 5.45, SBN 231; M 5.46, SBN 231; M 7.19, SBN 257; M 9.6, SBN 272; M 9.7, SBN 273; M 9.8, SBN 274.

I am therefore confident that, in the *Enquiry*, Hume endorses Generality, and that he frequently uses the term 'humanity' to refer to delicate sympathy, as understood in the *Treatise*. In the next chapter, I will largely return to his *Treatise*, to argue that Hume our consistent approval of justice as evidence for Generality and his thesis of delicate sympathy. It is only because of the uniformity of approbation, Hume argues, that we consistently approve of all motives to be just.

# 4. Hume on Justice

In Book 3 of his *Treatise*, Hume gives more consideration to what he calls 'artificial virtues' than to natural virtues (T 3.3.1.9, SBN 577). Artificial virtues are those that rely on conventions of language and behaviour which have arisen for non-moral reasons. Where we see that our adherence to any such convention is generally beneficial to society, we are very likely to come to view our adherence to it as morally obligatory.

The main such virtue which Hume discusses is 'justice', by which he means primarily a respect for the conventions concerning property rights and property transfer, such as the convention to pay back loans of money (T 3.2.1.1, SBN 477). By most interpretations of Hume's theory of moral judgements, there is much that is puzzling about his understanding of this virtue. This is particularly the case regarding his understanding of our motivation to act justly. Why do we feel obliged to act justly or honestly on those occasions when doing so will benefit nobody? Why should we, to use Hume's example, feel duty-bound to return a loan of money to a 'miser' or 'profligate debauchee' (T 3.2.1.13, SBN 482) when to do so will cause only harm to all concerned?

Hume clearly recognises that it is possible to feel such motivations. In the moral *Enquiry*, he suggests that only a 'sensible knave' would feel comfortable breaking the rules of justice, even in such situations (M 9.22, SBN 282-3). In the *Treatise*, he seems to believe that we are *all* motivated, at least to some degree, to perform all just acts, and that we all disapprove of those who do not. Yet why should we consider it vicious to break the rules on occasions where no one is harmed? Why are we not sensible knaves?

In a key passage, Hume focuses on two potential motives for justice. These are 'self-interest', which he calls the 'original motive to the establishment of justice', and 'a sympathy with public interest', which he describes as 'the source of the moral approbation, which attends that virtue' (T 3.2.2.24, SBN 499-500, italics omitted). These appear to be the *only* potential motives to perform just actions, according to Hume. Yet he allows that some just actions are both to the detriment of the actor's interests *and* such that 'the public is a real sufferer' (T 3.2.2.22, SBN 497). What, then, is Hume's explanation of the motivation to act justly in cases like this?

I will refer to this explanatory requirement as the problem of 'formally just' actions. A formally just action is one which is morally obligatory (insofar as it conforms to the rules of justice), but which the agent considering the action believes will cause no happiness to anyone. Indeed, formally just actions may be believed by the agent to cause only harm to all concerned. In this chapter, I will focus on Hume's *Treatise* account of justice to argue that existing interpretations of his treatment of formally just actions all face insurmountable difficulties. In contrast, by my interpretation of Hume's theory of approbation, as set out in Chapter 3, he has a simple and straightforward explanation of our approbation of formally just actions. Moreover, I will argue, he discusses formally just actions to begin his argument that the moral sentiments are all produced via delicate sympathy.

In §4.1, I address some of the general aspects of Hume's discussion of our approval of justice. In §4.2, I survey a range of interpretations of Hume's answer to the problem of formally just actions. I argue that they are all subject to insurmountable objections. In §4.3, I argue that Hume structured his *Treatise* so as to use the case of justice, with its prevalence of formally just actions, as evidence for Generality, as set out in §3.1.2, and for his thesis of delicate sympathy.

### 4.1. Moral obligation and justice

When Hume first discusses the virtue of justice, he focuses on what appears to be a fundamental problem with the notion of duty towards justice. Typically, when we act justly, we are motivated by a sense of duty. Indeed, Hume seems to claim that *whenever* we act justly, we do so purely from a sense of moral duty. We 'have naturally no real or universal motive for observing the laws of equity, but the very equity and merit of that observance' (T 3.2.1.17, SBN 483). However, Hume acknowledges that this claim appears to conflict with another proposition which he holds to be true, that 'no action can be equitable or meritorious, where it cannot arise from some separate motive' (T 3.2.1.17, SBN 483). By a 'separate motive', Hume means a motive distinct from the sense of duty. His point is that no action can be morally obligatory unless we have some motive other than a sense of duty to perform it.

This latter claim is a consequence of Hume's theory that we take an action to be virtuous because we feel approbation towards the motive or desire which led to that action. If I decide to be just, benevolent or in any way virtuous, my motivation cannot *only* be the dutiful desire to 'do the right thing', or to 'be virtuous'. This would be 'to reason in a circle', because, for Hume, 'virtuous' and 'approvable' are the same thing (T 3.2.1.4, SBN 478). If I decide to be virtuous, then I decide to do that which is approved, but what is approved is simply what is virtuous. Hume derives from this an 'undoubted maxim, *that no action can be virtuous, or morally good, unless there be in human nature some motive to produce it, distinct from the sense of its morality*' (T 3.2.1.7, SBN 479).

This antecedent, non-dutiful motive to perform any virtuous action is what Hume calls the 'original motive' for that virtue (T 3.2.1.13, SBN 482). It is easy to see what the original motives are for natural virtues. Hume gives the example of good parenting. We are generally not motivated to act in our child's interests because we think we ought to, but because we instinctively want to do so. Hume calls this instinct 'natural affection' and claims

that it is very clearly ‘a motive to the action distinct from a sense of duty’ (T 3.2.1.5, SBN 478). Hume allows that someone may decide to help their child from a sense of duty, but only because natural affection is already ‘common in human nature’ (T 3.2.1.8, SBN 479). If I normally delight in my child’s happiness, then on an occasion when I feel disinclined to act to please my child I will be pained by this disinclination. This may lead me to act ‘from a certain sense of duty’ (T 3.2.1.8, SBN 479). But this feeling of duty is, as Hume says with regard to benevolence, a ‘secondary consideration’ (T 3.2.1.6, SBN 478). If I did not realise that good parenting tends to cause happiness, then I would feel no duty to be a good parent.

The problem regarding justice is that it seems to have no original motive. I will call this the ‘original motive’ problem. Hume allows that there is a strong intuition that we are only ever motivated to be just from a sense of duty, and that there is no other obvious motive for just actions in many cases. Yet he also seems to claim that we could only feel duty-bound to be just if we also possess a distinct original motive towards justice. As Hume acknowledges, this appears to be ‘sophistry’ (T 3.2.1.17, SBN 483).

If we can understand how Hume resolves the original motive problem, such that he explains how we come to feel duty-bound to act justly in all cases, then this will resolve the problem of formally just actions. However, this is no easy matter. There is a formidable range of commentators, including Árdal (1966), Ayer (1980), Harrison (1981) and Gauthier (1992), who think that Hume’s account of our motivation for our own, and approval of other people’s, just actions is ultimately flawed. Baron (2001, 273) goes so far as to suggest that Hume tells a ‘noble lie’, and that he ultimately recognises that we have no moral reasons to perform many acts of justice, but refuses to state this, for fear of discouraging just actions in his readers.

There are however three potential options which Hume appears to have at his disposal, each of which will allow him to consistently categorise justice as a virtue and

explain our regard for formally just actions. First, he could point to his distinction between useful and agreeable virtues (T 3.3.1.30, SBN 590). Hume allows that some virtues, such as wit, please immediately, so that we find them agreeable even where we feel no sympathetic pleasures. I have argued that he thinks we experience *moral* approbation towards these virtues only via delicate sympathy, and I will return to this point in Chapter 6. Nevertheless, if acting justly is always immediately agreeable both in its performance and on witnessing it, then this might explain why we are motivated to be just or why we approve of just actions.[1] Second, Hume could argue that, despite appearances in some cases, we believe that all just actions *do* somehow directly produce sympathetic pleasures or prevent sympathetic pains. For example, on Stroud's (1977, 214) interpretation Hume thinks we believe that any omission of a just action would cause the institution of justice to 'collapse'. If this were the case, then one's motive to perform any just action would result, at least in part, from one's desire not to harm society at large. The third option for Hume is to argue that all just actions produce pleasures, or avoid pains, which are ultimately but *indirectly* derived from our sympathies with others. We will see that this is the most fruitful available option for Hume, and there are several interpretations of his theory which take this line.[2] In the next section, I will consider all three options, and argue that none of them are defensible as they currently stand.

[1] This is suggested by Krause (2004).

[2] This possibility has been suggested, within varying broader interpretations, by at least the following: Baier (1991); Besser-Jones (2006); Bricke (2000); Cohon (2010); Darwall (1993); Garrett (2007; 2015); and Stroud (1977).

### 4.2. Potential solutions to the problem of formally just actions

In what follows, I will consider each of the three options in turn. I will argue that only the final option is plausible, and that it is at its most plausible if we understand Hume to endorse Generality.

#### *4.2.1. Is justice an immediately agreeable virtue?*

Hume at least suggests that some character traits are considered virtuous because the actions in which they result are immediately agreeable, rather than because they cause happiness in those around us (T 3.3.1.27, SBN 589-90). While formally just actions are not useful to anyone, they may perhaps be intrinsically pleasing to perform and to witness. For example, Krause (2004, 644) suggests that we are taught that acting justly is inherently pleasing, by politicians and parents who 'make us regard the observance of the rules by which society is maintained "as worthy and honourable" in their own right, and their violation as "base and infamous"'. However, Krause also believes that Hume fails to articulate a persuasive account of the motive to justice, because the potential benefits of acting unjustly in such situations are so great, and because he has shown no way of assessing these potential pleasures against the pleasure of acting justly (Krause 2004, 645). An additional problem is that Hume frequently stresses the usefulness of justice, and asserts plainly that '*a* sympathy *with public interest is the source of the* moral approbation, *which attends that virtue*' (T 3.2.2.24, SBN 499).

#### *4.2.2. Do all just actions cause sympathetic pleasures or prevent sympathetic pains?*

Perhaps Hume believes that *all* just actions produce happiness to those around us, or at least prevent pains, despite occasional appearances to the contrary. In other words, perhaps he believes that there are *no* formally just actions. Hume could think this, for example, if he thinks that justice is a fragile enough institution to be threatened if any token just action is

disregarded, as if *any* token unjust action could cause the conventions of justice to be widely ignored. If this were true, then we would expect painful consequences, or at least a non-negligible probability of painful consequences, from any omission of a just action.

There are admittedly some indications that Hume holds such a view. He offers a quasi-historical account of how we came to have rules of justice, which suggests that the levels of cooperation we require as human beings in order to live comfortably on limited resources require us to develop such rules, initially from self-interest as a pragmatic measure. As Cohon (2010, 173) summarises it, we realise that we can 'invent rules attaching goods to individuals, and experience with the fragility of small societies teaches us that it is in our interest to conform to these rules'. During this quasi-historical account, Hume suggests that, although some just actions may cause unhappiness to all concerned:

> every individual person must find himself a gainer, on ballancing the account, since, without justice, society must immediately dissolve, and every one must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be suppos'd in society (T 3.2.2.22, SBN 497).

However, it does not seem plausible that anyone faced with the choice of repaying a miser will believe that society will collapse if they fail to do so. If that is Hume's argument, then he is providing only a forced choice, as Stroud (1977, 210) says, 'between a world in which everyone including himself is always just, and a world in which he and everyone else is always unjust', and so one where a just society collapses. I do not think we should consider this to be Hume's argument in the quoted passage. Here, Hume is not considering the reasons why we in a *modern* society would feel motivated to act justly. He is arguing that, in

primitive and small societies, the rules of justice are developed and obeyed from simple self-interest. It is only *later* that justice becomes approved of, and is therefore regarded as a virtue.

In the paragraph in question, Hume is still addressing the origins of the convention in a small society. His point is simply that, as soon as people in a small society design property rules, they must realise that the rules cannot allow of exception if they are to survive. In a small community not yet used to acting according to such rules, this seems plausible. It does not follow that in larger societies the same motive will hold. As we have seen, Hume acknowledges that we may firmly believe that breaching the rules of justice will have no negative effects but still disapprove of someone who does breach the rules, and be disinclined to breach them ourselves. Further, Hume has not yet discussed our *moral* motive to perform just actions at all, for he is discussing a point in time where that has not yet arisen. It is in the following paragraph that Hume asks '*[w]hy we annex the idea of virtue to justice, and of vice to injustice*' (T 3.2.2.23, SBN 498).

If Hume does not think that we (in our current large-society context) would expect society to collapse should we omit particular acts of justice, then there is no obvious reason to think that all just actions must cause more happiness than their omission. This brings us to the third possible solution for Hume, which is to claim that at least some acts of justice are motivated by the expectation of other kinds of pleasure than those derived directly from sympathy with others.

### *4.2.3. Do formally just actions cause pleasures not immediately derived via sympathy?*

If justice – and other artificial virtues – can be shown to have a motive which is not directly related to sympathy, then this might potentially explain our motivation to perform formally just actions. The most commonly suggested motive here is self-interest, or a minimally

extended version thereof, because Hume explicitly states that the rules of justice are initially developed and pursued out of mutual self-interest and a concern for the well-being of those dear to us: what Hume calls '*confin'd generosity*' (T 3.2.2.18, SBN 495). In this vein, Baier (1991), Bricke (2000) and Stroud (1977) interpret Hume as holding that we have self-interested reasons for universal compliance to the rules of justice. It is for this reason that Stroud, who thinks Hume's position is untenable, claims that Hume makes the implausible argument that society could collapse after any one token act of justice is omitted. Bricke (2000, 218) similarly takes Hume to believe that it is *always* in one's narrow interests to comply with the rules of justice. This thesis is so implausible that Bricke (2000, 216) is surprised at what he takes to be Hume's 'untrammelled confidence' in it.

The main concern with this general interpretation, however, is again that Hume only discusses purely self-interested motivations towards just actions as occurring in the *early stages* of the development of the motive to act justly. It is not until after he has concluded his account of the origins of justice that he considers why we are *morally* motivated to act justly. He is careful here to say that the reader must wait for the full answer to this, which is in 'the *third* part of this book' (T 3.2.2.23, SBN 498). There is therefore no reason to think that Hume attempts to explain our motivation to act justly purely on self-interest. The motive of self-interest is insufficiently moralised for it to be the motive Hume describes.

It may however be that just actions are pleasing because they are *generally* strongly associated with sympathetic pleasures, and that this is sufficient for us to approve even of rare token actions which will not produce such pleasures. This is, I think, the only plausible solution, but the question is how this association causes approbation in such situations. Interpretations along these lines often take Hume to have a complex story about our approval of justice. For example, Cohon (2010, 173) argues that there are three stages to the formation of an artificial virtue such as justice. The first stage is straightforward, as we create and

follow the rules from enlightened self-interest. Then there are two further stages as societies grow, which lead to this conformity becoming a moral practice. First, the mechanism of sympathy naturally makes everyone approve of other people's rule-following, because of the happiness which such rule-following causes and the misery which it prevents. Then there is a second artifice, this time performed by politicians and parents, which causes us to approve of honest characters as well as of honest actions. It is only after this second artifice that we are morally motivated to perform formally just actions, by the resulting 'enhanced moral sentiment' (Cohon 2010, 175).

By Cohon's (2010, 174n. 10) interpretation, Hume allows that we initially approve *purely* of just actions, regardless of motive, so long as they conform to the rules. Harris (2010) thinks Hume allows this for all cases of our approval of justice. However, Hume argues that, for any virtue, we *initially* approve only of motives: 'Actions are at first only consider'd as signs of motives' (T 3.2.1.8, SBN 479). Over time, the close association between virtuous motives and the actions which they typically produce is likely to cause us to approve of the actions themselves. Eventually, we may sometimes lack a token virtuous motive but nevertheless perform the associated action, 'merely out of regard to its moral obligation' (T 3.2.1.8, SBN 479). However, although Hume therefore allows that we *can* act justly because we feel approbation towards the idea of a just action, this can only occur once we habitually associate such actions with the relevant motive for justice: the fundamental source of this approbation.

Perhaps, however, we find justice generally pleasing because we approve of the *motive* to unquestioningly follow the rules, rather than of the rule-following actions themselves. This would be the case if the just person's motive is to regulate her conduct 'by rules she regards as authoritative', as Darwall (1993, 440) suggests. While he takes this to be Hume's account, he also argues that it is incompatible with Hume's theory of the will.

Darwall (1993, 423) understands this theory to stipulate that all just actions must be motivated by desires or aversions, caused by beliefs that pleasures will result or pains be avoided by so acting. However, Darwall (1993, 420) argues that the desire to follow a self-imposed rule would stem from no such belief, so that Hume's theory of justice can only be accepted if his 'official theory of will is jettisoned'.

Besser-Jones (2006) and Garrett (2007; 2015) both endorse Darwall's rule-regulation interpretation, but attempt to exonerate Hume from inconsistency. Besser-Jones (2006, 272) argues that the motive to regulate one's conduct by the rules of justice can be explained by reference to agents' desires to 'develop a good reputation and so become proud of their character'. Because we desire to maintain our reputation and character, both in the eyes of others and to ourselves, and because we are proud of our consistent adherence to justice, we are motivated to perform *all* just actions by the belief that this is necessary to preserve our pride and reputation. Therefore, Besser-Jones (2006, 262) argues that a rule-regulation account is compatible with Darwall's interpretation of Hume's 'theory of the will'.

Garrett (2007; 2015) similarly aims to defend the rule-regulation account against Darwall's charge of inconsistency. He argues that Hume allows for motivating passions to be caused by beliefs about the benefits and risks of general *policies*. When self-interest motivates people to create the convention of justice, it also 'gives rise to a new motive that could not have existed before: the desire and standing disposition to govern or *regulate* one's behavior by the rules of property' (Garrett 2015, 267, Garrett's emphasis). As we see the benefits of the entire scheme of justice, and as we similarly see that it is one which rests on everyone's observance of it, we come to approve of those who follow the rules unthinkingly, because we believe that such unthinking adherence to the rules is the best way of ensuring the efficiency of the scheme overall:

> In coming to regulate one's behavior in this way, one undertakes to refrain from weighing up the specific advantages and disadvantages of following the rules of property before acting in each individual case, for such weighing would often lead to violations (Garrett 2015, 267).

One worry here is that, while Garrett can explain why we might rarely *recognise* the harmful effects of certain acts of justice before we perform them, it is unclear that he can explain why we would feel duty-bound to perform a rule-following action in cases where we *do*, perhaps inadvertently, come to believe that only harmful consequences will result. However, there is a more serious problem for both Garrett's and Besser-Jones's interpretations, and indeed for all rule-regulation accounts. By any such account, Hume is understood to suggest a moral motive for justice which is quite unlike that for the natural virtues. It is essential to such interpretations that, once the rules of justice are established, just persons form a resolution to follow them unswervingly. This resolution is both the motive to perform just actions and the mental state we approve of in the just person. Of course, no such motive is required to explain our approval of the natural virtues: benevolence does not necessarily involve a *resolution* to consistently help others. Rather, any token desire to help others simply causes approbation. However, Hume is thought to require that we *resolve* to be consistently just, because this is deemed necessary for him to explain why we approve of *all* just actions, including formally just actions.

However, Hume argues that we can explain our approval of just actions in the same way that we explain our approval of *naturally* virtuous actions. Once the rules of justice have been established, 'the sense of morality in the observance of these rules follows *naturally*, and of itself' (T 3.2.6.11, SBN 533). Admittedly, as Besser-Jones stresses, Hume does also suggest a 'new artifice' by which 'the public instructions of politicians, and the private

education of parents, contribute to the giving us a sense of honour and duty in the strict regulation of our actions with regard to the properties of others' (T 3.2.6.11, SBN 533). However, he minimises the importance of this. He later emphasises, again, that once we have developed the relevant conventions and rules, justice is '*naturally* attended with a strong sentiment of morals, which can proceed from nothing but our sympathy with the interests of society' (T 3.3.1.12, SBN 579).

If Hume believes, not only that our *approval* of just actions develops 'naturally', but also that we are typically *motivated* by this approval, then he clearly means to explain our motive for justice in the same way that he explains our moral motivation to perform naturally virtuous actions. As desires to regulate oneself by rules are *not* required to explain the natural virtues, Hume cannot require any such motives for his explanation of justice.

We cannot simply assume that Hume thinks we are motivated to be just by our approval of justice. Árdal (1966) and Baier (1991) understand Hume to argue that, although we *approve* of justice because it serves the public good, we remain *motivated* to act justly purely from self-interest. Certainly, Hume denies that we act from any 'love of mankind, merely as such' (T 3.2.1.12, SBN 481). This is because no such passion exists: we do not *instinctively* love strangers. He also denies that we look 'so far as the public interest' when we act justly (T 3.2.1.11, SBN 481). We care little for the abstract idea of the public interest, and we do not perform just actions because we reason that they will benefit society. Any motive so formed would be too weak to overcome our self-interested desires, as justice frequently demands.

However, Hume does *not* conclude that our sole motive for justice is self-interest. He argues that our concern for the public good comes, not from instinct or reason, but from sympathy with the happiness of others (T 3.2.1.12, SBN 481-2). Once we are accustomed to justice, we are pleased – 'by sympathy' – that it benefits the public, and we therefore come to

approve of justice (T 3.3.1.9, SBN 577). Of course, this cannot explain why we *invented* justice, and so Hume argues that our original motive for justice was self-interest, whereas our approval developed later (T 3.2.2.24, SBN 499).

Hume believes that we, in our 'civiliz'd state', *are* motivated by a 'sense of duty and obligation' towards justice (T 3.2.1.9, SBN 479). We follow its rules *because* of our sense of duty, which gives us a 'regard to justice, and abhorrence of villainy and knavery' (T 3.2.1.9, SBN 479). Wherever we act from duty, we do so because our human nature possesses 'some distinct principles, which are capable of producing the action, and whose moral beauty renders the action meritorious' (T 3.2.1.8, SBN 479). By this, Hume means that to be motivated by duty is to be motivated by our approval of the idea of an action. Ideas of acting justly motivate via approbation, while ideas of acting unjustly repel us via disapprobation: 'a considerable motive to virtue', for those of us who are not knaves (M 9.23, SBN 283). Indeed, for non-knaves, these sentiments constitute a sense of moral obligation sufficient to motivate even formally just actions. We feel such sentiments because the *system* of justice tends to cause happiness, with which we sympathise (T 3.3.1.9, SBN 577). The question, therefore, is how the tendency of justice to cause happiness leads, *naturally* and via sympathy, to our approbation of, and so motivation for, *all* just acts.

In the next section I will argue that Hume believes that we feel morally obliged to perform any just action because we generally associate such actions with a 'tendency to the public good' (T 3.3.1.12, SBN 580). My interpretation certainly resembles the rule-regulation views of Darwall, Besser-Jones and, especially, Garrett, but there are two key differences. First, I do not take Hume to require that we must *resolve* to be consistently just in order to be consistently just. Second, I do not take Hume to argue that one's desire to perform a formally just action stems from any *belief* about beneficial consequences, whether of the token action or of justice more generally. Our awareness of justice's tendency to cause public good is

sufficient to produce moral desires to act justly, even where we believe that a token just action will produce only harm.

### 4.3. Instinct, artifice and Hume's arguments about justice

Consider the three-part structure of the third book in Hume's *Treatise*, 'Of Morals'. The first part is concerned to argue for the essential role of passions, over and above reason, in morality. The second – and largest – part is devoted to examining the artificial virtues. Here Hume argues, among other things, that we are motivated to perform and approve of formally just actions. The third part is ostensibly an account of the natural virtues, but in fact draws heavily on the prior discussion of the artificial virtues and discusses aspects of morality relevant to both kinds. Hume suggests several times in this section that his discussion of our approval of artificial virtues has important implications for his theory of our approval of natural virtues (T 3.3.1.10, SBN 577-8; T 3.3.1.12, SBN 579; T3.3.6.1, SBN 618).

I believe that Hume's aim in discussing the artificial virtues *before* the natural virtues is to demonstrate that a general association between a motive kind and the happiness of others is sufficient to cause approbation towards *any* token motive of that kind. Because Hume thinks this is easier to show for the artificial virtues than for the natural virtues, he discusses the artificial virtues first, and then applies this thesis to the natural virtues: 'We have happily attain'd experiments in the artificial virtues, where the tendency of qualities to the good of society, is the *sole* cause of our approbation, without any suspicion of the concurrence of another principle' (T 3.3.1.10, SBN 578). Hume could readily have said that the *good* caused to society by the artificial virtues, or our *belief* in this good, is the sole cause of our approbation towards them, had that been his meaning. Instead, he explicitly refers to the *tendency* of these qualities as being the cause.

In the final part of his book on morals, Hume describes the cause of approbation as

being the tendency to cause happiness (or that of disapprobation as the tendency to cause unhappiness) on no fewer than nineteen occasions.[3] As with his moral *Enquiry*, as we saw in Chapter 3, at no point in the *Treatise* does Hume suggest that we can only be morally motivated to perform an action where we hold certain *beliefs* about the consequences of doing so. Of course, this is not conclusive. No doubt one could read a sentence like 'qualities acquire our approbation, because of their tendency to the good of mankind' as meaning that we approve of qualities on token occasions when we *believe* they will cause some good (T 3.3.1.10, SBN 578). However, read literally, Hume is saying that we approve of quality types simply because they *tend* to cause good. With this in mind, the frequency and importance of the word 'tendency' in this final part of the book cannot be denied.

There is a good reason for Hume to begin arguing his case by discussing the artificial virtues, rather than the natural virtues. It is only with virtues like justice that we are likely to observe that some token actions are considered morally obligatory even where we do not expect them to cause happiness to anyone. As Hume argues at T 3.3.1.13 (SBN 580), one would rarely, if ever, be faced with a situation where one feels duty-bound to act kindly but where one knows that only unhappiness will result. It is simply not in the nature of kindness, unlike that of justice, for this to be a realistic possibility. Therefore, Hume begins his discussion of virtues by analysing the artificial virtues, because these provide crucial evidence for his thesis. If he can show that we approve of artificial virtues because they are *generally* associated with causing sympathetic pleasures, then he can claim the same form of general association as the cause of our approval of the natural virtues.

[3] T 3.3.1.9, SBN 577; T 3.3.1.10, SBN 578; T 3.3.1.11, SBN 579; T 3.3.1.12, SBN 580; T 3.3.1.13, SBN 580; T 3.3.1.14, SBN 580; T 3.3.1.19, SBN 584; T 3.3.1.23, SBN 586; T 3.3.1.25, SBN 588; T 3.3.1.27, SBN 589; T 3.3.1.28, SBN 590; T 3.3.2.15, SBN 601; T 3.3.3.3, SBN 604; T 3.3.4.5, SBN 610; T 3.3.4.11, SBN 612; T 3.3.5.1, SBN 614; T 3.3.6.1, SBN 618; T 3.3.6.4, SBN 620.

This is why Hume's discussion of justice focuses on the original motive problem, for it can only be this *original* motive which we come to associate with causing sympathetic pleasures, and which thereby causes us to approve of *all* just actions. The distinction between the natural and artificial virtues is, as we have seen, predicated on the former having instinctive 'original motives' which the latter lack. These original motives cause us to act in ways which, as we swiftly realise, tend to cause happiness in others.

Taylor (1988, 10) perspicaciously notes that "in his sketch of our pre-just moral psychology, the motives that Hume focuses on – sexual appetite, affection for children, limited benevolence, interest and resentment – are ones he characterizes in Book II as 'calm desires and tendencies'". These are all desires of the kind that, in §2.2, I called 'instinctive desires'. The instinctive desires are no more inherently moralised motives than is the desire to be just. However, some instinctive desires (and other non-artificial traits that are not desires, such as cheerfulness) tend to cause happiness to others, so that we come to approve of them. These are the natural virtues.

Hume explicitly describes the instinctive aspect of natural virtues in 'Of The Original Contract', where he says that people act in such ways because they are 'impelled by a natural instinct or immediate propensity, which operates on them, independent of all ideas of obligation, and of all views, either to public or private utility' (EMPL, 479). This is contrasted with artificial virtues, which are 'such as are not supported by any original instinct of nature, but are performed entirely from a sense of obligation, when we consider the necessities of human society, and the impossibility of supporting it, if these duties were neglected' (EMPL, 480).

Garrett (2007, 263) argues that Hume does not mean *moral* duty by 'obligation' here. He points to Hume's discussion of justice in the *Treatise*, where Hume suggests that we are initially motivated to follow the rules of justice by self-interest, and where he calls this

motivation a '*natural* obligation' (T 3.2.2.23, SBN 498). However, Hume does not mention the concept of natural obligation in *Of the Original Contract*. Equally, he would not have assumed that his reader would have read the *Treatise*, given that he had disowned it by the time of writing this essay. Furthermore, it is clear from Hume's description of natural virtues that he *is* referring to a specifically moral obligation. The instincts which cause us to perform natural virtues are contrasted with motives arising from ideas of obligation: an obligation which arises later when we come to pay natural virtues 'the just tribute of moral approbation and esteem' (EMPL, 479).

By the end of part two of Book 3 of the *Treatise*, then, Hume has divided all virtuous motives into motives for action types that we are instinctively inclined to perform and motives for action types that we are not instinctively inclined to perform. These latter virtues come into being because of human ingenuity and contrivance, and so are called 'artificial'. The natural and artificial virtues have this in common: in neither case do we initially perform the relevant actions with any idea of morality in mind. We are initially motivated to perform natural virtues by instinct, and we are initially motivated to conform to the rules of justice by our desires for the pleasures that will thereby be caused, or the pains that will thereby be avoided, for us and our loved ones. However, in both cases, because acting in these ways tend to produce happiness in others, all who witness such actions are frequently pleased by sympathy, and become accustomed to feeling pleased in this way. This leads us to approve of the motives to perform these actions – whether instinctive or (broadly) self-interested – and to categorise them as virtues. From this point on, we will feel obliged to perform such actions even when the original motive is absent. Eventually, we will approve of the actions which typically result from these motives as well, as Hume explains at T 3.2.1.8 (SBN 479). Our approval of justice will cause a calm desire to perform just actions even when we believe that *nobody* will be made happy by them.

Hume is now able to make his central claim about the causes of approbation. If we naturally come to approve of justice because the motives and actions involved *tend* to cause sympathetic pleasures in us, then such a tendency is sufficient to cause approval of any trait. Justice therefore provides good evidence for Generality. If Hume can show that we are motivated to perform formally just actions purely by delicate sympathy, approbation, and the desires that approbation can produce, then he can argue from this that all moral motivation is of precisely this kind.

We saw, in §3.3, that it is generally assumed that Hume allows only for believed ideas to motivate, but that this assumption is incorrect. If it were correct, then Hume would have to show that we *believe* that performing formally just actions is either certain or likely to cause pleasure or prevent pain, in order to explain why we are motivated to perform them. Hence, Garrett (2007, 274) takes Hume to argue that we 'come to believe' that it is in our interests to resolve to perform *all* just actions without considering the consequences. By Stroud's (1977, 214) interpretation, Hume thinks we have the 'false belief' that society will collapse if we omit *any* act of justice.

Hume requires no such belief. Once we associate just actions with causing happiness, we approve of *every* motive to act justly and disapprove of *every* motive to act unjustly. Equally, because these motives are strongly associated with the actions themselves, we come to approve of *every* just action and disapprove of *every* unjust action. Therefore, we feel duty-bound to act justly even where we believe that the action holds no prospect of pleasure for ourselves or others.

Not only does this interpretation resolve the problem of formally just actions, but it resolves the original motive problem as well. This resolution rests on the fact that Hume only claims, as Garrett (2007, 260) also notes, that we *naturally* have no 'real or universal motive for observing the laws of equity, but the very equity and merit of that observance' (T

3.2.1.17, SBN 483). While Hume requires that we have a motive to be just which is distinct from moral duty, this motive may occur *artificially*. I have argued against Garrett that Hume does not suggest that artifice leads us to develop an entirely new *kind* of motive, whereby we desire to avoid assessing the consequences of just actions. Rather, he simply means that the rules of justice were artificially created, and were initially followed from self-interest rather than by instinct. Once we are accustomed to follow the rules, we approve of the self-interested motivation to do so, in the same way that we approve of the instinctively arising motivations to help others or to act in one's children's interests.

I therefore understand Hume's treatment of justice to provide very strong evidence that he endorses Generality, and that he understands the moral sentiments to be uniform. In Chapter 5, I will argue that Hume understands *all* moral judgements to be moral sentiments, and that his is an emotivist theory of moral language.

# 5. Hume's Emotivism

Thus far, I have focused, as Hume does, on his understanding of the psychological causes of moral sentiments. I will continue to examine these causes in further depth, but I also want to begin considering how Hume understands moral language. What do we mean, according to Hume, when we sincerely call a person's motive 'virtuous' or 'vicious'? How does he understand the relations between such utterances and our sentiments of approbation and disapprobation?

Hume is undoubtedly, as Pigden (2007, 199) claims, 'widely regarded as the grandfather of emotivism and indeed of non-cognitivism in general'. Anyone, reading Hume for the first time, will be struck by language which seems to assert that any moral utterance is purely an expression of a moral sentiment (e.g. T 3.1.2.3, SBN 471). In this chapter, I will argue that Hume asserts precisely this. I will argue that he *is* an emotivist: he believes that any moral utterance derives its meaning purely from its being an expression of a moral sentiment.

There is strong, but not conclusive, evidence for this in Hume's moral *Enquiry*. In §3.1.2, we saw that Hume sees the moral sentiments as uniform: he believes that approbation, for example, responds in the same way towards all tokens of any one generally pleasing type of trait, regardless of how we are related to the person whose trait it is, or affected by the token trait. In §3.4, we saw that, in his *Enquiry*, Hume stresses that our verbal moral evaluations are *also* highly uniform: people typically call any act of benevolence 'virtuous', for example, no matter how the particular action affects them personally.

The uniformity of the moral sentiments makes them very useful to us, according to Hume, because it allows these sentiments to form the basis of a 'general system of blame or praise' (M 9.6, SBN 273). No other sentiments respond to characters or actions in any such

uniform manner. Hume certainly appears to argue that, for these reasons, we developed moral terms precisely to express only our uniform sentiments of approbation and disapprobation:

> The distinction, therefore, between these species of sentiment [i.e. the uniform and the variable] being so great and evident, language must soon be moulded upon it, and must invent a peculiar set of terms, in order to express those universal sentiments of censure or approbation, which arise from humanity, or from views of general usefulness and its contrary (M 9.8, SBN 274).

With this in mind, consider Hume's claim, that, when one person calls another '*vicious* or *odious* or *depraved*, he… expresses sentiments, in which, he expects, all his audience are to concur with him' (M 9.6, SBN 272). Hume certainly *appears* to mean by phrases like this, as indeed I think he does mean, that we express moral sentiments whenever we utter moral assertions.

However, Hume offers no arguments in his moral *Enquiry* to demonstrate that we cannot form and express moral *beliefs*, instead of or alongside our moral sentiments. Indeed, in the same passage just quoted, he suggests that someone who calls a man 'vicious' means 'to express, that this man possesses qualities, whose tendency is pernicious to society' (M 9.6, SBN 272). *This* appears to be an expression of a belief. On my reading, Hume simply means by this phrase that any disapprobation expressed via the term 'vicious' can only 'arise' – as suggested at M 9.8 – from the judger's sympathies with a quality that she associates with causing harm or unhappiness. However, Hume does not, unfortunately, make this explicit.

Nevertheless, Hume's language in his moral *Enquiry* strongly suggests that he is an emotivist. I will return to his *Enquiry* discussions of moral language in Chapter 6. There, I

will address Hume's thesis of a moral 'common point of view', which *appears* to suggest that many moral judgements are formed or corrected by reasoning (T 3.3.1.30, SBN 591; M 9.6, SBN 272). I will argue that this appearance is deceptive. However, it should already be apparent that there are reasons why readers of Hume might come to think that he allows for at least some moral utterances to express reasoned, moral beliefs.

Certainly, many influential readers of Hume – Pigden included – do *not* understand him to be an emotivist. This is largely because he has yet to be accredited with any convincing arguments for this thesis. In the face of this lacuna, those who do see Hume as arguing for emotivism generally think he does so unsuccessfully, perhaps because he has not sufficiently distinguished emotivism from other, similar but incompatible theses, such as subjectivism (e.g. Flew 1963; Harrison 1976; Mackie 1980). Some scholars, such as Cohon (2010), Garrett (2002; 2015), and Sayre-McCord (2008), argue that Hume *does* allow that some verbalised moral evaluations are expressions of moral beliefs. Others suggest that Hume is only concerned to argue that emotions have an important role in causing moral statements, rather than to account for the semantics of such statements, and that he therefore provides no considered semantic account (e.g. Árdal, 1966; Penelhum, 1975; Stroud, 1977). Of course, given that the conception of 'semantics' that we work with now is largely inherited from later philosophers, such as Frege, Russell, Tarski, and Davidson, this is hardly surprising. However, it is certainly the case that Hume says very little about meaning.

Nevertheless, I will argue that Hume's theory is an emotivist one. I do not intend to directly compare Hume's view with those of the 20th Century emotivists, such as Ayer, Carnap, or Stevenson.[1] As Sweigart (1964) argues, Hume's understanding of moral judgements is based on theories of belief, meaning, reasoning, and passions which are very

[1] I will, however, discuss Stevenson's views in detail in Part 2.

different from theirs. They certainly endorse no equivalent philosophical notions to Hume's 'vivacity' or 'sympathy' (Sweigart 1964, 232). Yet these are crucial to Hume's arguments.

If we are to understand Hume as an emotivist, we must do so in his terms. To this end, I will first argue that he follows Locke in understanding utterances to derive their meaning from the ideas (or impressions) that they express, so that his, largely implicit, theory of meaning is what we now call an 'ideational' one (e.g. Lowe 2006). I will then argue that, in his *Treatise*, Hume offers an important, albeit partially implicit, argument for *non-cognitivism*, by which I mean the thesis that no moral judgements are beliefs. Hume believes that *all* moral judgements are sentiments. From these interpretative claims, we can conclude that Hume is an emotivist: he believes that the meanings of *all* verbalised moral evaluations are derived purely from their being expressions of moral sentiments.

Hume's arguments for non-cognitivism are based in large part on his thesis of Generality, which (in the *Treatise*, at least) entails that moral sentiments are such as to arise *whenever* we contemplate morally relevant objects. Hume also, I will argue, endorses what I call a 'Vivacity Thesis': that wherever we have a present impression, we cannot hold a vivid idea in mind which differs from that impression only in its level of vivacity. Therefore, the presence of moral sentiments precludes any possibility of moral beliefs, because moral beliefs could only be less vivid copies of moral sentiments, and these cannot simultaneously exist.

Unfortunately, although I am confident that Hume *does* endorse the Vivacity Thesis, I am unsure whether he can consistently do so. We will see, in §.5.6, that the thesis relies on Hume's oft-stated claim that impressions and ideas differ from one another *only* in their different levels of vivacity. However, we have already seen, in §1.4, that Hume appears to implicitly allow for a further, fundamental difference: that all ideas are representative, as no impressions are. Perhaps Hume's arguments for emotivism cannot, therefore, be rendered

consistent with everything he says. Nevertheless, there can be no doubt that Hume claims, at several points in his *Treatise,* that impressions and ideas fundamentally differ only in their different levels of vivacity. Given this claim, I will argue, the *Treatise* provides an important argument for non-cognitivism. Moreover, we will see several other reasons to think that Hume denies the existence of moral beliefs. I will argue that we have sufficient reason to conclude that he is a non-cognitivist, and so an emotivist.

In §5.1, I address Hume's ideational theory of meaning. In §5.2, I consider textual evidence from the first part of Book 3 that Hume is a non-cognitivist. §5.3 discusses Cohon's and Garrett's arguments that Hume *does* allow for moral beliefs. In §5.4, I ask how Hume might potentially understand the causes and contents of such beliefs. In §5.5 I argue that, if he allows for the existence of moral beliefs, then, given his thesis of Generality, he would believe that any moral belief would be accompanied by a moral sentiment. Finally, in §5.6, I argue that Hume holds a Vivacity Thesis, which precludes the possibility of any moral belief existing alongside a moral sentiment directed towards the same object. I conclude that Hume does not allow for the existence of any moral beliefs, so that he is an emotivist.

### 5.1. Hume's theory of meaning

Locke argues that the purpose of language is to communicate our ideas to others. We utter words 'to make them stand as marks for the *Ideas*' within our minds (Locke 2008, 254). By this, Locke (2008, 254) does not mean that we refer to our ideas when we speak, but rather that we express these ideas: we speak so that our ideas 'might be made known to others, and the Thoughts of Men's Minds be conveyed from one to another'. If I say that something is 'red', then I do so to 'convey' that idea to others, so that they then think of it as red. As Locke (2008, 257) puts it, people utter words in order 'to bring out their *Ideas*, and lay them before the view of Others'.

Unlike Locke, Hume never provides anything like a theory of meaning. The closest he gets to discussing any such theory is to argue against Locke's theory of the meaning of abstract terms. As we saw in Chapter 2, he does so by arguing for his own theory of '*abstract* or *general* ideas' (T 1.1.7.1, SBN 17). Throughout this discussion, Hume clearly assumes that a general term derives its meaning from the general idea that it expresses, just as Locke claims. For example, Hume says that 'the mind' may consider the 'collection' of ideas within any one revival set, before it comprehends the general idea that it 'intends to express by the general term' (T 1.1.7.10, SBN 22).

Unlike Locke, Hume carefully distinguishes impressions from ideas. He sometimes, albeit rarely, explicitly allows that we express impressions as well as ideas. He worries that rival philosophers might 'express a hatred' of him and his work, for example (T 1.4.7.2, SBN 264). In Book 3, he argues that different 'sensations' of pleasure may 'be express'd by the same abstract term', provided that they sufficiently resemble one another to be recognised *as* pleasures (T 3.1.2.4, SBN 472).

Hume's theory of meaning can, therefore, at least allow that any utterance of a moral term can derive its meaning from its being an expression of a moral sentiment. He certainly believes that at least paradigmatic moral judgements are moral sentiments. From all that we have seen so far, it appears that Hume should understand any utterance of 'cruelty is wrong', for example, as an expression of disapprobation towards the general idea of cruelty, so that it derives its meaning from its being an expression of disapprobation towards the general idea of cruelty. I will argue that this appearance is correct, and that Hume believes that all verbal moral evaluations acquire their meaning by being expressions of moral sentiments. To do so, however, I must show that he denies that *any* moral judgements can be beliefs.

### 5.2. Hume (and Hutcheson) against the rationalists

Hume's most famous discussion of morality, in the first part of Book 3 of his *Treatise*, concerns the question of whether moral judgements can be formed purely via processes of reasoning. His first argument in this section is the 'Motivation Argument'.[2] Regimented somewhat, this is as follows:

1. morals… have an influence on the actions and affections
2. reason alone… can never have any such influence [on the actions and affections]
3. [so, morals] cannot be deriv'd from reason [alone] (T 3.1.1.6, SBN 457)

Although there are many interpretations of this argument, the most influential one has Hume arguing that (1) moral judgments are motivating mental states, that (2) reasoned beliefs do not motivate on their own, so that (3) moral judgments cannot be reasoned beliefs.[3] However, as Snare (1975) argues, Hume's rationalist opponents can simply respond that one of the premises is only typically true, but not necessarily so, which allows that some or all moral judgments may be beliefs. Their best argument is presumably to say that, although experience demonstrates that most beliefs are motivationally 'inert', *some* beliefs can motivate on their own, most notably moral beliefs. Unless Hume can demonstrate that beliefs *necessarily* cannot motivate in this way, his argument simply begs the question. As it is, his argument appears to offer little to persuade anyone already committed to rationalism.

Why would Hume begin his book on morals with such an unpersuasive argument?

[2] I borrow this name from Cohon 1997.

[3] Commentators who endorse something like this interpretation of Hume's argument, such that it is an argument about the nature of moral judgements, include the following: Bricke (2000); Harrison (1976); Mackie (1980); Penelhum (1975); Snare (1975); and Stroud (1977).

Some, such as Cohon (2010) and Garrett (2002), believe that we have misinterpreted Hume, and that his argument is stronger than it seems. Cohon (2010, 81-90) claims that Hume means to argue only that the process by which we make a moral judgement is not purely a process of reasoning, so that he does not argue for anything directly about the nature of moral judgements themselves. Garrett similarly claims that Hume's focus is not on the nature of individual moral judgments, but rather on the 'general question of whether the *origin* of the capacity to make moral distinctions depends only on *reason* or on something else (such as distinctively moral impressions)' (Garrett 2002, 193, Garrett's emphasis).

However, Hume certainly seems to deny that *any* moral judgements are beliefs. Consider, for example, his claim towards the end of this section, that 'when you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it' (T 3.1.1.26, SBN 469). Perhaps, as Garrett (2002, 202) argues, we should not take Hume to be offering an analysis of the meaning of a moral sentence or utterance here. Nevertheless, it is hard not to read this as the claim that *every* token moral judgement is wholly a matter of sentiment, and not of belief. Hume's final conclusion in this section is 'that the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceiv'd by reason' (T 3.1.1.27, SBN 470). The scope of this claim is less clear, but Hume certainly appears to deny that reason *ever* produces moral perceptions. This, again, seems to preclude that any moral judgements can be beliefs.

Whether or not Hume's conclusions in T 3.1.1 clearly assert that all moral judgements are matters of sentiment rather than of reasoned belief, his summary of his conclusions at the start of T 3.1.2 seem unequivocal:

> To have the sense of virtue, is nothing but to feel a satisfaction of a

> particular kind from the contemplation of a character. The very feeling constitutes our praise or admiration. We go no farther; nor do we enquire into the cause of the satisfaction. We do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous. The case is the same as in our judgments concerning all kinds of beauty, and tastes, and sensations. Our approbation is imply'd in the immediate pleasure they convey to us. (T 3.1.2.3, SBN 471)

Hume never mentions any cases where a moral judgement might be a belief, and he appears to be strongly arguing, throughout T 3.1.1, that *no* moral judgements are beliefs.

Moreover, as Stroud (1977, 173) notes, Hume has already argued, in Book 2, that it is a contingent but nevertheless well-evidenced truth that no reasoned beliefs are motivationally efficacious on their own, and that all motivation requires passions of some kind. Given this, the Motivation Argument appears at least reasonably well founded, even on the standard interpretation. Hume may well see it as a good starting point from which to develop his moral sentimentalism, by which all moral judgements are passions. If we allow that the Motivation Argument is based on Hume's theories of ideas and of the passions, then it at least demonstrates that moral judgements are much more *likely* to be passions than beliefs, and so sets the scene well for Hume to begin developing his sentimentalist theory.

I think this is how we should read the Motivation Argument. I think that T 3.1.1 is intended to show that, at least given the theories of reasoning, passions, and beliefs for which Hume has argued in Books 1 and 2, moral judgements are not beliefs. Unfortunately, however, it is not easy to understand how to read Hume in this section. This is largely because, as Cohon (2010, 79) stresses, Hume carefully avoids any explicit discussion of

moral judgements throughout it. Indeed, his language is frequently, and confusingly, ambiguous. He never clearly states what, in his view, it is that cannot be produced purely by reasoning. Instead, he makes vague claims, such as that 'morals' cannot be 'deriv'd from reason (T 3.1.1.6, SBN 457), that the 'rules of morality… are not conclusions of our reason' (T 3.1.1.6, SBN 457), or that '[m]oral distinctions… are not the offspring of reason' (T 3.1.1.10, SBN 458).

Why does Hume not employ the terms that he has so carefully defined over Books 1 and 2 of his *Treatise*, and claim that we never form moral beliefs, and that what appear to be moral judgements are in fact calm passions? Why, too, does he rely on arguments that would convince only readers who already agree with his own conceptions of reasoning, belief, and passion? I suggest that the answer to these questions can be found in Hume's desire to persuade Hutcheson, and his followers, of the merits of Hume's own sentimentalist theory.

### *5.2.1 Hutcheson's moral sense theory*

Following the arguments of Moore (1994), as well as of Gill (2010, 213) and Harris (2015, 123-124), I believe that T 3.1.1 was a late addition to Hume's *Treatise* discussion of moral judgements. Moore (1994, 39) argues – persuasively, to my mind – that Hume wrote T 3.1 with the primary aim of influencing Hutcheson and his followers to endorse his own sentimentalism over Hutcheson's 'moral sense' theory. I suggest that Hume's language throughout most of T 3.1.1 is ambiguous because he uses this section to attempt a delicate balance, between convincing Hutcheson and his followers of the similarities between his and Hutcheson's theories, and laying the groundwork for his own, very different, theory.

Hutcheson's *Illustrations on the Moral Sense* were published only twelve years before Book 3 of Hume's *Treatise*. In the *Illustrations*, Hutcheson argues for the existence of a 'moral sense', which allows us to form ideas of virtue and vice, much as other senses allow

us to form ideas of secondary qualities, like colours or sounds. Hutcheson (1971, 163) allows that these kinds of 'sensible ideas' are 'only perceptions in our minds'. He argues that 'approbation cannot be supposed an image of any thing external, more than the pleasures, of harmony, of taste, of smell', but that this does not 'diminish' the 'reality' of our idea of virtue (Hutcheson 1971, 164).

Hutcheson also provides several arguments against moral rationalism which are strikingly similar to Hume's arguments in T 3.1.1. Darwall (1997, 73) argues that 'the main lines of [Hume's] account, as well as significant details, derive directly from Hutcheson'. Gill (2010, 213) similarly argues that Hume's arguments against rationalism 'do not advance significantly' beyond Hutcheson's, although Moore (1994, 39) suggests that Hume uses more sceptical arguments than Hutcheson's. Certainly, Hume appears to have been heavily influenced by Hutcheson in T 3.1.1.

Hutcheson sees moral philosophy as a contest between two camps, both of which are internally consistent, but only one of which can be correct. One camp is Hutcheson's own, which claims that 'we have not only self-love, but benevolent affections also towards others' (Hutcheson 1971, 118). The other is an 'Epicurean' camp, which includes Hobbes and certain 'Christian moralists' who claim that all actions are ultimately motivated by self-interest (Hutcheson 1971, 117). Hutcheson clearly sees it as his duty to argue against moral scepticism and egoism. He also denies that the rationalists form a genuine alternative to these two camps, because he believes they have simply confused the moral sense with the dictates of reason. They are, he thinks, merely misguided members of his own camp. Therefore, the goal of the *Illustrations* is to examine the rationalists' theories and 'to explain more fully how the moral sense alleged to be in mankind must be presupposed even in these schemes' (Hutcheson 1971, 119).

For Schneewind (1998), the key threat to morality in Hutcheson's eyes is the cynical,

egoist philosophy of Mandeville. Mandeville argues that moral behaviour was unknown in the state of nature, and that it is no part of human nature as such. We only came to think in moral terms, claims Mandeville, when certain clever people realised that they could manipulate us by flattering our self-esteem, and by claiming that following their preferred rules made us honourable. To defeat this view, as Schneewind (1998, 336) claims, Hutcheson argues for 'three connected theses':

> He must show that there is an idea of virtue that is different in nature from any idea concerned with self-interest. He must show that benevolence alone is what we approve of morally. And he must show that benevolence or unselfish concern for others is a natural and effective human motive (Schneewind 1998, 336).

With these theses in mind, we can return to T 3.1, to examine how Hume treats the same topics.

*5.2.2. Hume's response to Hutcheson's moral sense theory*

In T 3.1.1, Hume argues, as Hutcheson had done, that moral rationalists are mistaken, because we *feel* virtue and vice, rather than use reasoning to form beliefs about virtue and vice. As we have seen, his argumentative strategy closely follows Hutcheson's own. At the end of this section, Hume suggests, much as Hutcheson has done, that vice and virtue 'may be compar'd to sounds, colours, heat and cold, which, according to modern philosophy, are not qualities in objects, but perceptions in the mind' (T 3.1.1.26, SBN 468-9). This is not a claim that Hume makes elsewhere in his *Treatise*, or in his moral *Enquiry*, and it is far from clear that he can mean anything substantive by it. Indeed, he has previously argued, in T

1.4.4, that there is no meaningful distinction between primary and secondary qualities. *All* qualities that we appear to perceive in objects are 'perceptions in the mind', according to Hume.[4]

Hume then reassures his reader that his is neither a cynical nor sceptical account of morality:

> Nothing can be more real, or concern us more, than our own sentiments of pleasure and uneasiness; and if these be favourable to virtue, and unfavourable to vice, no more can be requisite to the regulation of our conduct and behaviour (T 3.1.1.26, SBN 469).

In this way, T 3.1.1 demonstrates precisely those areas where Hume's theory of moral judgements aligns with Hutcheson's, while carefully avoiding any areas of difference between the two. Hume says nothing *here* of benevolence, or of natural motives, about which he has very different views to Hutcheson.

One difference between the two theories is purely terminological, albeit with important consequences. Unlike Hutcheson, who closely follows Locke in his use of the term 'idea', Hume carefully distinguishes ideas from impressions. Indeed, in his first *Enquiry*, Hume complains that 'the word *idea*, seems to be commonly taken in a very loose sense, by Locke and others; for standing in for any of our perceptions, our sensations and passions, as well as our thoughts' (E 2.9n1.1, SBN 22). Hutcheson is, almost certainly, one of the 'others' about whom Hume complains.

[4] Blackburn (1993b, 275) similarly argues that Hume merely uses this phrase from 'piety to Hutcheson', and to stress that moral judgements are not reasoned beliefs.

Hutcheson (1971, 164) claims that we form 'ideas of virtue and vice' by sensing them, via certain pleasurable and painful feelings. In Hume's terminology, Hutcheson would presumably understand moral approval as an 'impression of sensation' (T 2.1.1.1, SBN 275). However, Hume carefully distinguishes 'the passions' from 'the impressions of the senses' (T 2.1.1.1, SBN 275). Hutcheson's 'ideas' of virtue and vice are more similar to Hume's 'passions' than to his 'impressions of sensation' *or* 'ideas'. Hutcheson might not be unhappy to hear that Hume sees them as impressions, not ideas. However, if Hume were to explicitly claim that moral judgements are calm passions, then he would be arguing *against* Hutcheson.

This is, I suggest, the reason for Hume's otherwise perplexingly ambiguous language throughout T 3.1.1. Hume and Hutcheson agree *only* to the extent that they both understand moral judgements as pleasurable and painful feelings, which cannot be produced purely by reasoning. This, and no more than this, is what Hume argues for in T 3.1.1. He employs his technical language only to argue that moral judgements are feelings, nothing more.

Hume starts T 3.1.1 by carefully setting out his question in general terms: "*Whether 'tis by means of our* ideas *or* impressions *we distinguish betwixt vice and virtue, and pronounce an action blameable or praise-worthy*?" (T 3.1.1.3, SBN 456). To 'pronounce an action blameable or praise-worthy' is, surely, to make what would normally be considered a moral 'judgement'. Hume's question is therefore that of whether what we take to be moral judgements are impressions or ideas. His answer, at least for cases where we 'pronounce any action or character to be vicious', is also clear: we do so only by means of 'a feeling or sentiment of blame' (T 3.1.1.26, SBN 469). We saw, in Chapter 1, that Hume does not use the term 'sentiment' in any technical sense: it is sometimes a synonym for taste and it sometimes means something like 'opinion'. Given this, T 3.1.1 seems very carefully worded to appeal to Hutcheson.

Hume also carefully refrains from mentioning moral 'beliefs' or 'judgements'

throughout T 3.1.1. Indeed, he never once uses the specific phrases ‘moral judgment’ or ‘moral judgments’ in Book 3. He fairly often talks of other kinds of ‘judgment’ and ‘judgments’ in Book 3, however. In the *Treatise*, Hume takes a judgement to be a belief: one formed at the conclusion of some process of reasoning. At least, he suggests this meaning of the term at T 1.3.13.19 (SBN 153). He argues that what we take to be moral judgements are, in fact, impressions, rather than ideas. Therefore, they cannot be beliefs, and so they cannot be judgements in Hume’s sense of the term.[5] I think we should take him entirely literally when he claims that morality ‘is more properly felt than judg'd of’ (T 3.1.2.1, SBN 470).

This claim appears at the start of T 3.1.2. Hume then argues, along very Hutchesonian lines, as Gill (2009b, 575) observes, that virtue feels pleasurable to contemplate and that vice feels painful to contemplate. After this, he asks ‘*in general*, concerning this pain or pleasure, that distinguishes moral good and evil, *From what principles is it derived, and whence does it arise in the human mind?*’ His response to this question, from T 3.1.2.6 to T 3.1.2.10, is mainly to reject various Hutchesonian theses: notably, his thesis of benevolence and his thesis of moral approval as a natural motive. As Moore (1994) observes, Hume had previously attacked these theses in a letter to Hutcheson, sent just before he completed his final draft of Book 3, in 1739.

In the letter, Hume claims that ‘Were Benevolence the only Virtue, no Characters cou’d be mixt, but wou’d depend entirely on their Degrees of Benevolence’ (HL1, 34). In T 3.1.2.6 (SBN 473), he argues that ‘the number of our duties is, in a manner, infinite’. Hume claims that this makes it very unlikely that we have an instinctive moral sense, for there are simply too many virtues for us to instinctively respond to. Human nature, according to Hume,

[5] Nevertheless, just as Hume often uses common language rather than his own terminology, I will continue to discuss Hume’s theory of moral judgements.

is unlikely to have developed so many similar but distinct instincts, and so we should 'abridge these primary impulses, and find some more general principles, upon which all our notions of morals are founded' (T 3.1.2.6, SBN 473). Throughout Book 3, Hume argues for many kinds of virtue, of both 'natural' and 'artificial' kinds, that we come to value via the psychological processes of general rules and delicate sympathy, *not* via brute instinct.

Hume also argues, in T 3.1.2, against the claim that 'the character of natural and unnatural can ever, in any sense, mark the boundaries of vice and virtue' (T 3.1.2.10, SBN 475). He asks the Hutchesonian question of whether the principles of virtue are natural, and he responds that 'our answer to this question depends upon the definition of the word, Nature, than which there is none more ambiguous and equivocal' (T 3.1.2.7, SBN 474). He then argues that, whatever definition we choose, we cannot make sense of the idea that virtue is natural and vice unnatural. Indeed, he goes so far as to state that 'nothing can be more unphilosophical than those systems, which assert, that virtue is the same with what is natural, and vice with what is unnatural' (T 3.1.2.10, SBN 475). Moreover, in his 1739 letter to Hutcheson, Hume says bluntly that 'I cannot agree to your Sense of Natural. Tis founded on final Causes; which is a Consideration, that appears to me pretty uncertain & unphilosophical' (HL1, 33).

For Hume, benevolence *is* a natural virtue, but he disagrees with Hutcheson's conception of nature. For Hutcheson, 'natural' is a justificatory term. We naturally, and so correctly, approve only of 'impartially benevolent intentions' (Gill 1996, 27-28). However, associations of ideas may cause us to approve of other things which are not truly virtuous, and such approval will be unnatural, false and distasteful. As Gill (1996, 24) puts it, '[s]omething is truly beautiful or virtuous, Hutcheson believes, if and only if it would cause or occasion the appropriate type of pleasure in someone whose constitution is in its original pre-associative state.' Hutcheson never fully explains why this should be, but Gill (1996, 28)

suggests that his idea of nature sits within a 'theological world-view, and [that] the viability of that world-view was not something Hutcheson thought to submit to philosophical scrutiny.' All 'true explanations ultimately end at God' and so, when we can no longer explain why we feel something, we attribute the feeling to a God-given sense (Gill 1996, 28). For Hutcheson, this both sufficiently explains the feeling in question and justifies it. Hume clearly does not subscribe to any such view.

Hume's final paragraph of this section begins:

> Thus we are still brought back to our first position, that virtue is distinguished by the pleasure, and vice by the pain, that any action, sentiment or character gives us by the mere view and contemplation. This decision is very commodious; because it reduces us to this simple question, *Why any action or sentiment upon the general view or survey, gives a certain satisfaction or uneasiness*, in order to shew the origin of its moral rectitude or depravity, without looking for any incomprehensible relations and qualities, which never did exist in nature, nor even in our imagination, by any clear and distinct conception (T 3.1.2.11, SBN 475-6).

This all strongly suggests two things. First, that Hume believes that what we call moral 'judgements' are *all*, properly speaking, feelings of 'satisfaction' or 'uneasiness', rather than beliefs or ideas. And, second, that his arguments for this lie mainly beyond the first part of Book 3. T 3.1.1 appears to have been intended to show where Hume and Hutcheson agree: in their claims that moral judgements cannot be produced purely via reasoning, and that they must be feelings or sentiments. T 3.1.2 appears to have been intended to argue against Hutcheson, by showing that we cannot appeal to notions of naturalness or benevolence to

either explain or justify our moral sentiments in anything like the way that Hutcheson suggests. At the end of this section, having explicitly argued against Hutcheson's views regarding benevolence and our natural dispositions to be virtuous, and having implicitly ruled out any of Hutcheson's theological claims, Hume is left only with the question of how we *should* understand our moral sentiments.

Hume certainly never suggests that any moral judgements might be beliefs: 'Our decisions concerning moral rectitude and depravity are evidently perceptions; and as all perceptions are either impressions or ideas, the exclusion of the one is a convincing argument for the other' (T 3.1.2.1, SBN 470). Moral judgments, or more properly, moral 'decisions', are not ideas, and so not beliefs. Even if T 3.1 provides no compelling argument for non-cognitivism, it provides very good reasons to think that Hume *endorses* non-cognitivism.

Hume's substantive thesis, that all moral judgements are calm passions, is to come, in T 3.2 and T 3.3. We have examined the core tenets of this already. We will see, in §5.6, that Hume's official theories of belief and of the causes of the moral sentiments entail that we *cannot* have moral beliefs alongside moral sentiments. First, however, we must ask what a moral belief *would* look like, if Hume were to allow that they exist.

### 5.3. Arguments against Hume's non-cognitivism

Cohon (2010) and Garrett (2002) suggest two means by which Hume seems to allow moral beliefs to be formed. Cohon argues that we can form moral beliefs *without* inference, and I discuss this in in §5.3.1. In §5.3.2, I discuss Garrett's claim that we can use reason to form moral beliefs.

Both Cohon and Garrett see an important parallel between Hume's accounts of *sense* impressions and *moral* impressions. They suggest that Hume sees virtue as importantly similar to a sensory property like redness. We understand that an object looks red because we

see it looking so, and we therefore form beliefs about unperceived red objects by having vivid ideas of them as having this unanalysable property. Similarly, they argue, Hume believes that we understand a character to be virtuous because it *feels* a certain way, so we can form beliefs about absent virtuous characters by forming vivid ideas of them with *this* unanalysable property.

In §5.4, I will argue that Hume does not allow this parallel. To perceive something as red is just to have an impression of it *as* red, because Hume denies that colour and form arise from distinct impressions. Instead, we see impressions of colour 'dispos'd in a certain form' (T 1.1.7.18, SBN 25). My belief that a box is red just is a vivid idea of it as a box with that property I call redness. However, to take something to be virtuous is to experience approbation as *caused by* that object. I argue, therefore, that any belief that an object is virtuous could only be a belief that it is *such as to cause* approbation; not a belief that it has a certain quality of virtue. First, however, I will consider Cohon's and Garrett's arguments in greater detail.

### *5.3.1. Humean moral memories*

For Cohon (2010, 138), 'Hume's position is that ordinarily I acquire moral beliefs as the result of feeling a trait's goodness or evil – sensing the moral property directly – and then forming an idea-copy of my moral sentiment'. She argues for a direct comparison with beliefs about colour; 'normally I obtain my beliefs about what colors things are in response to a sensory experience' (Cohon 2010, 104). If I see a red box, then Hume's account seems to suggest that I will form an idea-copy of its colour which, being vivid, will be a belief that the box is red. Surely, argues Cohon, Hume must allow that we similarly form beliefs directly from our *moral* sentiments. If I feel approbation towards someone, then presumably I will likewise form a vivid idea-copy of this approbation; a moral belief.

However, Hume does *not* allow that we form beliefs directly from impressions. He claims that we may form a vivid or a non-vivid idea-copy – or, more likely, both – from any impression, but neither kind will be a belief. To form a vivid idea-copy is to have a memory. These can never be beliefs, because all beliefs exist within the faculty of the imagination, which is distinct from that of the memory (T 1.1.3.1, SBN 8). All ideas formed within the imagination are 'perfect', non-vivid ideas (T 1.1.3.1, SBN 8). Although we can later enliven these by causal reasoning, they begin as merely imagined ideas. Any beliefs about whether the objects which we perceive are 'really' of the colour they appear are formed by reasoning about the external world, as Hume argues in T 1.4.2. Therefore, we cannot *directly* form a belief from any impression, whether of a colour or a moral sentiment.

However, even if Cohon is describing what Hume would understand as moral memories rather than moral beliefs, she suggests a non-emotivist interpretation of his account of moral judgements. If I feel disapprobation towards Rousseau, then I judge him to be vicious. If I remember that I felt disapprobation towards Rousseau, then I have a vivid idea of his viciousness. In expressing this memory, I express my judgement of his viciousness, without any sentiment occurring. Therefore, I will have to demonstrate that Hume does not allow that *memories* could be moral judgements.

### *5.3.2 Humean moral beliefs*

Garrett (2002) argues, like Cohon, that Hume must allow that we can form ideas from moral impressions, just as we do from sense impressions. He claims that we could therefore *reason* with these ideas, and thereby 'formulate propositional judgements or beliefs to the effect that a particular person *is virtuous*' (Garrett 2002, 197, Garrett's emphasis). We are, Garrett suggests, particularly likely to engage in such reasoning where we are too distant from people for their characters to readily cause moral sentiments in us. He argues that denying that we

can reason about the morality of distant people would be like claiming that we 'could never *infer* that a box was red or square – say, as the result of someone else's testimony that it was' (Garrett 2002, 200).

Garrett's (2002, 200) example is that someone could 'infer from the testimony of the *New York Times* that Mother Teresa is virtuous'. Of course, by any account, such a judgement will be caused to some extent by reasoning. There will be inferences to the effect that Mother Teresa is a real person, that the writer is reliable, and so on. However, Garrett's claim is that Hume would allow that someone could form the judgement that Mother Teresa is virtuous by reason *alone* in this case. If so, then any judgement so formed would be purely a moral belief.

This is possible, according to Garrett, because of Hume's account of abstract ideas. We form our revival set of 'virtue' by grouping together all and only those ideas of 'personal characteristics that produce immediate moral approbation' (Garrett, 2002, 197). Garrett argues that, once we have categorised a sufficient number of ideas in this way, we can form beliefs about appropriate new members of the set *without* feeling approbation, by noticing resemblances between these traits and ones previously classed as 'virtuous'. The reader of the *New York Times* understands that Mother Teresa has traits of a kind which she usually calls virtuous and so, without any sentiment necessarily occurring, she applies the term 'virtue' to the idea of Mother Teresa's character. She therefore believes that Mother Teresa is virtuous.

If moral reasoning of this kind is possible by Hume's account, as it currently seems to be, then what, precisely, is being believed about Mother Teresa at its conclusion? Garrett thinks the belief is that Mother Teresa's character has an unanalysable property; that what is believed 'can only be expressed by saying that she is *virtuous*' (Garrett 2002, 202, Garrett's emphasis). In §5.4, I argue, against Garrett, that any such belief could only be that Mother Teresa's character is such as to cause approbation.

### 5.4. The vivid idea of virtue

Any belief about virtue must be a *vivid idea* of virtue. In the moral *Enquiry*, Hume 'defines virtue to be *whatever mental action or quality gives to a spectator the pleasing sentiment of approbation*' (M App. 1.10, SBN 289). His account in the *Treatise* is consistent with this definition (see particularly T 3.3.1.30, SBN 591). Therefore, any idea of a virtuous character trait just *is* a complex idea of both a mental action or quality and an associated sentiment of approbation. It is because we observe that some complex ideas resemble one another in this way that we form the revival set of 'virtue'. The idea of Mother Teresa's character could only be included, and so sincerely called 'virtuous', if it resembles other ideas of virtuous characters *by bringing to mind a further idea of approbation*. Therefore, the belief that Mother Teresa is virtuous *is* the belief that her character is such as to cause approbation.

Garret can resist this conclusion, because he claims that Hume allows *two* ways in which one can form the abstract idea of virtue, only one of which requires an idea of approbation. Garrett (2002, 107) argues that Hume defines 'personal merit' as the possession of traits which display 'usefulness or agreeableness to the possessor or others', and that, because all and only such mental qualities cause approbation, Hume considers this to be a definition of personal merit *and* of virtue. If so, then ideas of personal merit would also be ideas of virtue, but ones involving no ideas of approbation.

However, Hume never *claims* that 'the possession of useful or agreeable mental qualities' is a definition of virtue. Garrett (2002, 107) only points to one passage to support this interpretation, where Hume says that every useful or agreeable trait 'communicates a pleasure to the spectator', and is *then* deemed virtuous (M 9.12, SBN 277). Yet this merely supports Hume's explicit definition of virtuous traits as ones which causes approbation. It is, for Hume, merely a contingent fact that all and only useful or agreeable traits cause

approbation. The idea of *virtue* is of traits which cause approbation, although the same traits are also understood, by causal reasoning, to be useful and agreeable ones. Nevertheless, the abstract ideas of virtue and personal merit remain distinct; one could *conceive* of a useless, disagreeable virtue, simply by imagining such a trait causing approbation.

Therefore, any belief that a character is virtuous can only be, for Hume, a vivid idea of it as a potential cause of approbation. If the reader believes that Mother Teresa's character is *virtuous* – rather than just that the journalist at the *New York Times* approves of her – then she must believe that Mother Teresa's character is one which she, the reader, associates with causing approbation. She need not realise that this is *all* that there is to virtue, of course. Equally, she might not realise that Mother Teresa's character is such as to cause a *sentiment* of approbation, because Hume allows that we frequently mistake calm sentiments for reasoned judgements (T 2.3.3.8, SBN 417). Nevertheless, whatever she takes approbation to be, she would certainly understand that a virtuous trait is one which *would* cause her to approve, if she were in a position to witness it.

Any such belief could only be formed by causal reasoning. To believe that Mother Teresa is virtuous is to have a vivid idea of approbation as caused by her character. This idea is believed because of the vivacity transferred to it, from the reader's impressions of the newspaper, via the idea of Mother Teresa's character. If Hume *does* allow for moral reasoning, then any moral belief so produced will be a belief that a person's character is such as to cause approbation or disapprobation in the right circumstances. We will see the importance of this in §5.6.

In summary, Cohon's and Garrett's accounts suggest that we can form moral judgements in the absence of sentiment, just as we can form causal beliefs and memories about the colours of unperceived objects. All such non-sentimental moral judgements will be either memories or causal beliefs.

### 5.5. Implications of Generality

We have seen that the most plausible interpretations of Hume as allowing for non-sentimental moral judgements rely on an analogy between moral sentiments and sense impressions. This analogy suggests an argument that, because we form ideas from both kinds of impression, we can therefore believe or remember either kind of idea in the absence of the relevant impressions. Just as ideas derived from sense impressions can be judgements about unperceived objects, so too can ideas derived from moral sentiments be judgements about virtue and vice. However, Generality, as discussed in 3.1.2, shows a crucial difference in the causes of each kind of impression. This has important implications for any interpretation of Hume's account of moral reasoning.

Whereas sense impressions cannot be caused by ideas, a sentiment of approbation can – and will – be caused by any idea of a useful or agreeable character, action or sentiment. Indeed, Hume's theories of secondary impressions and of custom entail that, even if I were told only that someone was virtuous, with no detail about their character beyond this, I would form a general idea of their possessing useful or agreeable character traits, and would therefore feel approbation. Hume argues that we all 'implicitly' maintain a belief that virtuous traits are ones which cause happiness to others (M 9.2, SBN 269). This is presumably because the causal relation between happiness and approbation is so strong that we are able to make the inference from one to the other without even considering the matter. Where we have frequently experienced causal relations, 'the understanding or imagination can draw inferences from past experience, without reflecting on it' (T 1.3.8.13, SBN 104). Any persuasive testimony to the effect that someone is virtuous will cause me to infer that they possess character traits which are naturally fitted to be useful or agreeable. As *all* perceptions of such traits cause approbation, this belief will cause approbation.

On Hume's account, therefore, reasoning about virtue is in all cases redundant, because *whenever* we have an idea of a useful or agreeable trait, we will judge it as virtuous simply by experiencing approbation. After all, why *reason* with those ideas derived from our sense impressions? Only because sense impressions of the relevant kinds are frequently absent when we consider their causes and effects. In contrast, Hume thinks we need never reason about the causes of our passions because, *wherever* we consider the usual causes of a passion, we will feel some degree of the passion in question.

This is similar to a point which Blackburn (1993b) makes about the disanalogies between colour perceptions and moral feelings, regarding the corrections we may wish to make to our judgements about these. I may be confused about the morality of an action because my selfish interests are affecting my judgement, just as I may be confused about the colour of an object because of bad lighting. However, Blackburn notes that the differences are crucial:

> In the latter case we have only a judgement about what we would perceive were the light different. We do not have another colour perception alongside whatever we are seeing at the time. Whereas in the ethical case, we do have a genuinely moral sentiment emerging from the process, another original existence to put alongside whatever initial sentiment self-love generated. (Blackburn 1993b, 275)

We can transpose this thinking to the question of reasoning about the virtue of others. If I am told by a reliable witness that they passed a letter box, then I can infer that it is red, but I cannot *see* its redness. However, if a reliable witness tells me that a person helps others, then I have no need to *infer* that this person is virtuous, for the idea of helping others causes

approbation. Equally, if they persuade me simply that a person is virtuous, then I will form a belief that their character is useful or agreeable. Again, this belief will cause approbation.

Generality entails that every moral judgement involves an occurrence of approbation or disapprobation. This is because we experience at least one such sentiment whenever *any* morally relevant idea comes to mind. Equally, it demonstrates that reasoning about the morality of unperceived objects is simply not necessary in the way that reasoning about colour often is. However, this is not sufficient to demonstrate Hume's emotivism. Hume's claim is that we *never* form reasoned moral judgements; that every judgement of virtue is 'nothing but' a sentiment (T 3.1.2.3, SBN 471). Even given Generality, we *could* hold a belief about Mother Teresa's virtue whilst simultaneously feeling approbation towards her. Indeed, Garrett (2002, 198) suggests that we may have a moral belief and a 'corresponding moral feeling' simultaneously. To show that Hume denies this possibility, we must consider his account of vivacity once more. We will see that Hume's understanding of beliefs and of impressions compel him to deny that we can ever feel moral sentiments and moral beliefs simultaneously.

### 5.6. The Vivacity Thesis

The Vivacity Thesis states that, wherever we have a present impression, we cannot hold a vivid idea in mind which differs from that impression only in its level of vivacity. Therefore, we cannot form a *belief* about a moral sentiment which, aside from its level of vivacity, is identical to a present moral sentiment.

Admittedly, Hume never explicitly states the Vivacity Thesis. It is however entailed by a small number of key claims in Book 1. First, as we saw in Chapter 1, Hume claims that any idea of X is identical to an impression of X, except that it is less vivid. Hume claims this of impressions and ideas of a particular shade of red, and he argues that 'the case is the same

with all our simple impressions and ideas' (T 1.1.1.5, SBN 3). Hume allows that there is more than one 'kind of approbation', but only where different kinds of useful and agreeable traits cause variation in the feeling of approbation (T 3.3.4.2, SBN 608). Therefore, Hume's official view is that any believed idea of approbation as caused by an object would be identical to, except less vivid than, an impression of approbation as caused by the same object.

Second, Hume says that 'every distinct perception, which enters into the composition of the mind, is a distinct existence, and is different, and distinguishable, and separable from every other perception, either contemporary or successive' (T 1.4.6.16, SBN 259). If I believe that a box looks red, at the same time that I see it looks red, then I must be able to distinguish between my impression of red and my vivid idea of red. I would only be able to distinguish these perceptions by their differing levels of vivacity, as they are otherwise identical.

Third, Hume claims that we cannot doubt the existence of our perceptions, because they are 'immediately present to us by consciousness' (T 1.4.2.47, SBN 212). We *can* confuse one kind of perception for another, as when we confuse a calm passion for a reasoned belief, but we cannot doubt that the perception exists. If I believe that a box looks red, at the same time as I see that it looks red, then I must be able to identify that I have *two* perceptions of redness, which are distinguishable only by their differing levels of vivacity.

With this in mind, recall from Chapter 1 that all impressions are maximally vivid, whereas no beliefs are. If I see a box and have an impression of red, then I am *certain* that I see red, according to Hume; I know I *cannot* be mistaken here. If I believe that a box looks red, then, by definition, I am *uncertain* that I would see red when looking at it; I recognise *some* possibility of error. To be clear, the belief in question is not that the box is *always* or *really* red. It is simply a belief that the box looks red. If I were to maintain this belief at the same time as I look at the box and see its redness, then I would be in a state of both certainty

and uncertainty about what is otherwise the *exact same judgement* about the box's redness. Given that this is impossible, Hume must hold the Vivacity Thesis.

For sense impressions, this has only a limited effect. I cannot form a belief that a box looks red while I see that it does, simply because at that time I *know* that it looks red. However, if the box is removed from my sight, then the impression of red is gone, and I can remember the box's redness, and form beliefs about its redness. For moral impressions, however, the case is different.

We saw in §5.4 that any belief that Mother Teresa is virtuous would be a vivid, complex idea of her character and of approbation as caused by it; a belief that her character *would* cause approbation in the right circumstances. Generality entails that, if I read in the *New York Times* that Mother Teresa possesses any useful or agreeable traits, then I will immediately feel approbation. Equally, if it reports simply that she is virtuous, then I will habitually form an idea of her as possessing useful or agreeable traits, which will cause approbation. If I do not *believe* that Mother Teresa possesses the traits which I read about, then I will merely experience approbation towards the general ideas of the traits; I will consider them virtuous but not applicable to Mother Teresa. If, however, I infer that Mother Teresa does have such traits, then this approbation will be caused by what I believe to be *her* traits. It will be a judgement of her virtue.

This impression of approbation as caused by the idea of Mother Teresa's character will be identical to any believed idea of such approbation, except in its level of vivacity. The Vivacity Thesis entails that I cannot simultaneously experience these two perceptions. Therefore, I can never simultaneously experience the vivid idea of Mother Teresa's useful or agreeable traits and the vivid idea of a consequent approbation, because the former idea immediately *causes* approbation. As any belief about Mother Teresa's virtue could only consist of two such vivid ideas held together, I cannot form the belief that she is virtuous. In

other words, I cannot merely *believe* that approbation is caused by a character while approbation is caused by it, any more than I can merely *believe* that I see a box as red while I see its redness. Similar arguments apply to *all* potential beliefs about virtue or vice. Therefore, Generality and the Vivacity Thesis together demonstrate that we cannot *ever* hold beliefs about virtue or vice.

Equally, moral judgements cannot consist of memories of moral properties, as Cohon's interpretation suggests. In fact, we cannot *ever* form memory ideas of the virtuousness or viciousness of characters. Like beliefs, memory ideas are vivid, but less so than the otherwise identical impressions from which they are derived. Therefore, the Vivacity Thesis demonstrates that any memory idea of the viciousness of Rousseau's ingratitude can only be formed once I have ceased to feel disapprobation. Generality demonstrates that, if I remember Rousseau's ingratitude, I will once again feel disapprobation.

Therefore, Hume's is an emotivist theory: if I sincerely call Rousseau 'vicious', then I can only be expressing a sentiment of disapprobation towards Rousseau, and my utterance can only derive its meaning from its being an expression of a sentiment of disapprobation towards Rousseau.

It is notable that, while the Vivacity Thesis disallows *moral* beliefs (and other beliefs about matters of taste), it allows beliefs about violent passions, because we can distinguish a calm passion from a violent one of the same kind. If I imagine a bear in my garden, for example, then Hume's theory of passions suggests I will feel some, presumably *very* calm, fear. I will also believe that I would feel a *more violent* fear if there *was* a bear in my garden. This belief is therefore distinguishable from my impression of fear, so both can exist simultaneously. However, to approve of a future action or distant character is to have an impression as calm as any approbation one might believe one *would* experience. This

impression differs from the belief *only* in its level of vivacity, so the two cannot exist simultaneously.

As a final point, Hume's theory allows that we can entertain *non-believed* ideas of virtue and vice, and even believed ideas of approbation and disapprobation. If I believe that 'the monks perversely believed that silence was a virtue', then I form a vivid idea of their approbation as caused by silence; I believe that *they* take silence to be virtuous. Further, I can imagine approving of silence, so that I imagine that silence is a virtue. In both cases, because I do *not* approve of silence, I experience no simultaneous impression of approbation, and I form no belief about the virtue of silence. However, if I come to approve of silence, then I will experience approbation, so that, according to the Vivacity Thesis, I will be unable to form a vivid idea of my *own* approbation as caused by silence. I will be unable to believe in the virtue of silence.

Hume's ideational theory of language, coupled with his thesis that moral judgements are all sentiments rather than beliefs, entails an emotivist theory of moral language. Admittedly, Hume's Vivacity Thesis is founded on a claim – that impressions and ideas fundamentally differ only in their different levels of vivacity – that Hume appears unable to consistently adhere to, as we saw in Chapter 1. Nevertheless, he resolutely adheres to this claim throughout his *Treatise*. Moreover, he appears adamant in his rejection of the notion of moral beliefs, in both his *Treatise* and his moral *Enquiry*.

We have one further, and important, aspect of Hume's theory of moral judgements still to discuss: his thesis of a common or general point of view. As we will see, it is often believed that this thesis requires that some or all moral judgements are produced or corrected by reason. I will argue, in Chapter 6, that this belief is a mistaken one.

## 6. The Common Point of View

In both his *Treatise* and his moral *Enquiry,* Hume argues that we would frequently disagree with one another when assessing other people's characters, were it not for the fact that, in relevant cases, we adopt a 'common point of view' (T 3.3.1.30, SBN 591; M 9.6, SBN 272). Adopting this viewpoint provides us with a 'general inalterable standard, by which we may approve or disapprove of characters and manners' (T 3.3.3.2, SBN 603; see also M 5.42, SBN 229). In the *Treatise*, Hume describes this as the process of fixing 'on some *steady* and *general* points of view', in which we consistently 'place' ourselves when we assess people's characters (T 3.3.1.15, SBN 581-2). In this chapter, I will ask how we should understand Hume's thesis of a common or general point of view. (I will generally use the phrase 'common point of view'.)

I have argued that Hume endorses the thesis that, in §3.1.2, I called Generality. Generality allows Hume to explain cases of virtue in rags *and* our consistent approval of the artificial virtues, without requiring any additions or caveats to his theory of the causes of moral judgements. According to Generality, our moral sentiments are *uniform*: they respond in the same way towards all token character traits of any one type, regardless of how we are related to the person whose trait it is, or how we are affected by the particular effects of the token trait. In this chapter, we will see that Generality cannot be reconciled with Hume's common point of view thesis, as it is typically understood. However, I will argue for a new interpretation of this thesis, which *can* be reconciled with Generality. This argument will conclude my discussion of Hume.

There is much debate about Hume's meaning in his discussion of the common point of view, but it is generally agreed that he is arguing for something like the following view: The sympathetic basis of our moral sentiments makes them variable in ways that our

considered moral judgements are not. For example, we sympathise more with people nearby than with those further away, and so we experience stronger moral sentiments towards character traits that affect people around us than towards similar character traits in distant lands. However, common experience shows that we generally evaluate similar traits by using similar moral terms, regardless of these kinds of variations: our *verbalised* moral evaluations are highly uniform. For example, we consistently call anyone who helps others 'good' or 'virtuous', no matter where they are or how we may be related to them. The reason for the uniformity of our verbalised moral evaluations is that we all recognise that our moral sentiments are variable, and we correct for these variations by undertaking an imaginative exercise: one that involves the adoption of a common point of view. Our motive to correct our judgements in this way is often understood to be, as Baier (1991, 181) has it, an awareness of the 'biases to which we know felt sympathy to be subject', along with a desire to correct for these when moralising. Once we have adopted the common point of view, our verbal evaluations are uniform because they are expressions of either suitably corrected sentiments or of our beliefs about 'how we *would* feel' if our sentiments were not 'influenced by our particular perspectives' (Radcliffe 1994, 43, Radcliffe's emphasis).

This is, of course, a highly simplified and generalised account. Nevertheless, most scholars agree that Hume holds some version of the thesis outlined above.[1] Call it the 'Correction Thesis'. This holds, roughly, that adopting the common point of view involves disregarding immediate and variable (i.e., *non*-uniform) moral sentiments in favour of more

[1] Proponents of this general view include, but are not limited to, the following: Abramson (1999); Árdal (1966); Baier (1991); Blackburn (1998); Bricke (2000); Brown (1994; 2001); Cohon (2010); Darwall (1994); Garrett (2002; 2015); Gill (2009b; 2010); Hearn (1976); Korsgaard (1999); Mackie (1980); Mercer (1972); Radcliffe (1994); Sayre-McCord (1994); Stroud (1977); and Taylor (2002; 2015).

considered, uniform moral judgements.[2] These corrected judgments may be understood as sentiments or, as we saw in chapter 5, as beliefs.

Hume is typically thought to argue that we perform this kind of correction via a sympathetic exercise, in which we imagine the actual or typical effects of a person's character on her 'narrow circle', sympathise with their responses, and thereby judge the character more accurately or impartially than we could do otherwise (T 3.3.3.2, SBN 602).[3] In some passages, such as T 3.3.1.15 (SBN 581-2) and T 3.3.1.20 (SBN 584-5), Hume stresses the similarities between our assessments of people's characters and those of beauty. This leads some readers, such as Garrett (2015, 121) and Sayre-McCord (1994), to compare Hume's discussion of the 'general inalterable standard' of morality with that of the 'true standard' of taste by which a competent critic judges art (EMPL 240). By these interpretations, adopting the common point of view is, roughly, a moral version of the process which Hume thinks a good art critic undertakes, when she imagines herself in the point of view of the intended audience of an artwork, sympathises with the audience's reaction, and judges the work purely by the resulting, refined sentiments.

---

[2] I use the term 'roughly' because my arguments will also oppose variations on this general interpretation, such as those by which any truly *moral* sentiments are only experienced following a corrective 'imaginative exercise', as Abramson (1999, 343) argues. I see no place for any such imaginative exercise.

[3] For interpretations by which we correct our initial moral sentiments by imaginatively sympathising with a person's 'narrow circle', see, for example: Brown (1994, 24); Darwall (1994, 71); Gill (2010, 253-254); Korsgaard (1999, 3); Radcliffe (1994, 42); Sayre-McCord (1994, 219). Taylor (2002; 2015) argues that, although Hume claims in his *Treatise* that the common point of view involves sympathising with a person's narrow circle, he does not hold this view in his moral *Enquiry*.

In this chapter, I will argue that Hume never argues for the Correction Thesis. This is important, because the Correction Thesis is typically seen as an integral and prominent aspect of Hume's theory of the formation of moral judgement. Moreover, many commentators claim that the thesis entails that at least some moral judgements are beliefs, closely analogous to reasoned beliefs, or formed mainly via processes of reasoning.[4] Absent the Correction Thesis, there appears to be no plausible place for reflective reasoning in Hume's account of the causes of our moral judgements.

This alone suggests one reason to think that Hume never argues for the Correction Thesis.[5] In Chapter 5, we saw his uncompromising anti-rationalism throughout T 3.1.1, which appears to have been written *after* his discussion of the common point of view. Hume's arguments in this section contain no mention of the adoption of a common point of view. If he saw any important role for reasoning in this process, then we should expect him to discuss it in T 3.1.1.

We must therefore ask what Hume *does* mean to argue for in his discussions of the common point of view, if not for the Correction Thesis. I will argue that he claims that, whenever we evaluate motives and characters, we habitually correct for the presence of our variable, *non-moral* sentiments, by expressing only our uniform moral sentiments. To adopt the common point of view *just is* to express only our automatically produced, uncorrected, and uniform moral sentiments in our evaluations of character. The process of adopting the

---

[4] Examples include: Aiken (1979); Baier (1991, 179-80); Cohon (2010, 150); Garrett (2002, 203); Hearn (1976, 63); Mercer (1972, 69); Norton (1982, 95); Sayre-McCord (2008, 312); Stroud (1977, 192).

[5] Korsgaard (1999), Mackie (1980), and Stroud (1977) argue for other problems for this thesis. Taylor (2002; 2015) argues for problems for Hume's *Treatise* account of the thesis.

common point of view requires no alteration of our sympathies or moral sentiments, and no imaginative exercise.

Recall, from §2.2, that Hume believes that we may 'turn our view' to any one idea in several different ways (T 1.1.7.18, SBN 25). In this way, the idea can be a more or less generalised one: an idea of a token character trait can function either as an idea of a person's particular trait, or as the general notion of that trait (such that it is thought of simply as a token of some general type), or as the abstract idea of that general type of trait. I believe that Hume has, at least sometimes, a similar meaning of 'view' in mind when he discusses a 'common point of view'. To adopt the common point of view is, in both the *Treatise* and the moral *Enquiry*, to express only the uniform sentiments that are caused, via delicate sympathy, when we view our ideas of traits as general notions, as where we think of a motive of benevolence merely *as* a motive of benevolence.

However, there are important differences between the two works. As Taylor (2002; 2015) argues, Hume's account of moral judgement in his *Enquiry* significantly improves on that in his *Treatise* in its greater focus on moral language. I argue that, in the *Treatise*, Hume claims that we recognise that our moral sentiments are more uniform than all our other sentiments directed towards character traits, which compels us to try and *alter* our non-moral sentiments accordingly, or at least to talk as if we have done so. We call a distant benefactor and a nearby benefactor equally 'good' because we mistakenly believe that we are rationally obliged to feel the same violent pleasures towards both benefactors. However, we can only successfully talk as if we feel equally about them by expressing only our uniform sentiments. In the *Enquiry,* Hume argues for the simpler thesis that we recognise that our moral sentiments are more uniform than all our other sentiments directed towards character traits, and that we find this pleasing and socially useful, because it allows for a high degree of

convergence in our evaluations of character. Hume argues that, for this reason, we developed moral language to express only these sentiments.

In §6.1, I address Hume's *Treatise* discussion of a common point of view. I focus on the problem that this discussion is intended to resolve: that of satisfactorily explaining why our verbal assessments of character are less variable than most of the passions which result from contemplating people's characters. I call this the 'uniformity problem'. I argue that Hume's response to it is heavily influenced by a Hobbesian theory of value, which in turn influences his claims about how and why we express only our moral sentiments when we evaluate characters. In §6.2, I argue that, in the *Enquiry*, to take up the common point of view just is to express our uniform sentiments via moral language.

### 6.1. The common point of view in the *Treatise*

Hume first discusses the common point of view in T 3.3.1, just after he has completed his account of the artificial virtues, and during the course of his arguments for Generality. He introduces the notion of 'delicate sympathy' at T 3.3.1.8 (SBN 576-7). He then argues that the case of justice has demonstrated that our 'sentiment of morals… can proceed from nothing but our sympathy with the interests of society' (T 3.3.1.12, SBN 580). He reminds us that some 'particular' acts of justice are 'not beneficial to society', and that we only approve of justice because of its general 'tendency' to benefit society, before arguing that we should 'ascribe… the same cause to the approbation of [the natural virtues]' (T 3.3.1.13, SBN 580).

Hume then addresses two potential objections to his argument so far. One stems from cases of virtue in rags, as discussed in §3.1. The other stems from the obvious variability of our typical sympathetic pains and pleasures. For example, Hume believes that our sympathies with our countryfolk are stronger than those with foreigners, so that we feel greater sympathetic pleasures from the happiness of an English person than from the happiness of a

Chinese person. However, we 'give the same approbation to the same moral qualities in *China* as in *England*' (T 3.3.1.14, SBN 581). If I believe that a Chinese person is as benevolent as an English one, then I will call the two characters equally 'virtuous'. Yet, Hume suggests, if my approval is caused by my sympathetic pleasures, and if I feel more pleasure from a beneficial outcome in England than from a similar one in China, then we would expect me to express greater 'esteem' towards the English benefactor than towards the Chinese benefactor (T 3.3.1.14, SBN 581). Following Cohon (2010, 131), I call this the 'variability objection'.

Whatever Hume means by '*steady* and *general* points of view', he introduces this phrase in response to the variability objection (T 3.3.1.15, SBN 581-2). And, however we understand the details of this objection, it is clearly directed towards Hume's suggestion that our moral judgements are caused via sympathy. Hume believes that sympathy is 'very variable', and his worry is that 'it may be thought, that our sentiments of morals must admit of all the same variations' (T 3.3.1.14, SBN 580-1).

As we have seen, Hume is typically understood to accept that our immediate, unreflective 'sentiments of morals' *are* variable, and to argue that the Correction Thesis resolves any worries we might have about this. By any such interpretation, the variability objection is understood to pose something like the following question: Why are our considered moral judgements uniform, given that our immediate and unreflective moral judgements are sentiments, which vary as our sympathies vary? However, if this *is* the relevant question, then Hume should reject it, rather than attempt to answer it. This is because, as we have also seen, he is in the midst of arguing that our immediate, unreflective moral sentiments are uniform.

Generality entails that, if I contemplate distant benevolent motives, or even benevolent motives that lead to painful consequences for me personally, then I will

immediately experience strong approbation, precisely as I do in those cases of virtue in rags where I approve of futile benevolent motives. There is no need for me to adopt an imagined viewpoint to achieve this. Given Generality, adopting an imagined viewpoint could only affect my moral sentiments if it resulted in me reclassifying the motive under consideration. For example, I might initially take someone to be benevolent but, by reflecting on how their peers might see them, decide that they are acting from self-interest. However, *this* kind of correction cannot be what Hume argues for in response to the variability objection, because it cannot explain why an English person judges a Chinese benefactor to be as virtuous as an English benefactor.

At this point in the *Treatise*, Hume has not yet completed his arguments that general rules and delicate sympathy cause uniform moral sentiments. He does this just after he discusses the variability objection, during his response to the 'virtue and rags' objection, at T 3.3.1.20 (SBN 585). Nevertheless, the thesis that he sets out to defend from the variability objection is carefully formulated to be compatible with Generality. It is the thesis that, wherever a character 'has a *tendency* to the good of mankind, we are pleas'd with it, and approve of it; because it presents the *lively idea* of pleasure; which idea affects us by sympathy, and is itself a kind of pleasure' (*T* 3.3.1.14, SBN 580, my emphasis). This clearly allows that any token act of a typically useful or agreeable type may cause approbation in the manner discussed in Chapter 3: via a process whereby general rules produce a quasi-belief about happiness, with which we sympathise via delicate sympathy.

Unfortunately, Hume's language seems less than careful where he argues for and defends this thesis. He frequently refers to violent passions as simply 'passions' or 'sentiments', and to calm passions by terms like 'calm and general principles' (T 3.3.1.18, SBN 584). This is at least partly because he believes that we commonly mistake our calm passions for reasoned judgments, so that we wrongly believe that our only passions are

violent ones. In T 3.3.1.18 (SBN 583-4), Hume briefly acknowledges that he is following our common language in this regard (see also T 2.1.1.3, SBN 276).

Hume begins his response to the variability objection by rejecting any notion that moral judgements are 'deriv'd from reason' (T 3.3.1.15, SBN 581). He stipulates instead that moral judgements are derived from a 'moral taste' and 'certain sentiments of pleasure and disgust' (T 3.3.1.15, SBN 581). These seem like synonymous terms, but I think they are not. Hume is responding to the suggestion that our moral 'esteem… proceeds not from sympathy' (T 3.3.1.14, SBN 581). I propose that he means that *some* non-moral 'sentiments of pleasure and disgust' must cause our moral taste, and that he is defending the thesis that these non-moral sentiments are *sympathetic* pleasures and pains. He concedes that these 'sentiments, whence-ever they are deriv'd, must vary, according to the distance or contiguity of the objects' (T 3.3.1.15, SBN 581). His point here, I suggest, is that whether or not the kinds of non-moral sentiments that cause our moral sentiments are sympathetic ones, they are clearly passions of kinds that typically vary in their violence, depending on the distance or contiguity of their objects. He freely allows that sympathetic pleasures are typically variable, although he will soon argue that all moral sentiments are caused by uniform, calm, 'delicate' sympathetic pains or pleasures (as we saw in §3.1.2).

Hume claims that, when we judge a historical character trait alongside a friend's trait of the same kind, we find a 'variation of the sentiment, without a variation of the esteem' (T 3.3.1.15, SBN 581). Although he appears to mean by this that our initial moral sentiments vary, unlike our more considered moral judgements, I think he means that our (non-delicate) sympathetic pains and pleasures vary, whereas our moral language does not. His point is that we cannot feel the same 'lively' pleasure from the virtuous motive of an ancient Greek that we feel towards that of a friend, despite esteeming both equally (T 3.3.1.15, SBN 581). Of course, according to Hume's official use of the term in the *Treatise*, *all* feelings of pleasure

are equally and maximally 'lively', or vivid, because they are all impressions (T 1.1.1.12, SBN 7). Therefore, Hume presumably means that the passions that vary with sympathy are, to a greater or lesser extent, *violent* ones. These cannot be moral sentiments, which are all calm passions, and thus easily confused with reasoned beliefs. Hume confirms this point a few paragraphs later, at T 3.3.1.18 (SBN 583).

Hume is asking why it is that, even where we feel more violent sympathetic pleasure towards a friend than towards a historical character, our *language* remains uniform: 'I do not say, that I esteem the one more than the other' (T 3.3.1.15, SBN 581). His immediate answer is that *any* 'system' of moral judgements that conforms to his stipulated requirements will face this kind of worry: a proponent of any such system must reconcile the uniformity of our moral language with the variability of our typical passions, regardless of whether or not 'sympathy' plays any important role within that system (T 3.3.1.14, SBN 581). Given the sympathetic basis of his own system, this suggests that he feels the need to explain why our verbalised moral evaluations are uniform, despite the fact that we clearly feel variable sympathetic pleasures and pains towards character traits. This, in turn, suggests – I think rightly – that the variability objection poses only the following question: if moral language expresses sentiments that are caused by sympathetic pains and pleasures, then why is this language uniform, when our sympathetic pains and pleasures are typically variable? Why, for example, do we call *all* benevolent people 'virtuous', even where they are too distant for us to feel any violent sympathetic pleasure from their actions?

Hume's theory of delicate sympathy, soon to be fully explained, will provide the core of his answer to this question, by showing how approbation occurs uniformly towards all motives of any one useful or agreeable type, regardless of any variations in our non-delicate sympathetic responses. However, I believe that Hume also aims to respond to a further, and generally unappreciated, worry for his theory, which I call the 'uniformity problem'. This is,

in part, the problem of explaining why those sentiments that we call 'moral' play greater roles in our social lives than any of the other sentiments that we experience towards people's characters.

### *6.1.1. The uniformity problem*

To help explain the uniformity problem, I should first note that Hume begins T 3.3.1. by summarising his claim, argued for in Book 2 and in T 3.1.1, that all motivation and evaluation is ultimately grounded in our sentiments and passions (T 3.3.1.2, SBN 574). This is not a new position, of course, and Hume and his readers will be aware of Hobbes's claim that '*Good* and *evil* are names that signify our appetites and aversions' (L 15.40). We will see reasons to think that Hume was influenced in T 3.3.1. by Hobbes's discussion surrounding this claim.

The uniformity problem demands a satisfactory answer to the following question: Why are our verbal evaluations of character uniform, when most of the passions that occur when we contemplate characters are variable? Generality cannot fully resolve *this* problem, for Generality only pertains to approbation and disapprobation, and these are merely two sentiments among many. Even if I experience the same approbation towards a benefactor in China as towards a similar one in England, Hume believes that I also feel more violent, non-moral pleasures towards the English benefactor than towards the Chinese benefactor, due to my stronger sympathies with English people. Given Hume's theory of value, it seems that he should expect me to say that, from *my* point of view, the English benefactor is better than the Chinese one, just because I feel more pleasure from contemplating the former than the latter. Of course, we do not talk like this, and Hume needs to explain why not.

Even if Hume is not directly influenced by Hobbes, it is helpful to consider Hobbes's own moral theory, to see the kind of view that Hume needs to reject: one by which we simply

call anything that pleases us ‘good’. In his own, brief, treatment of ‘moral philosophy’, Hobbes argues that ‘private appetite is the measure of good and evil’ (L 15.40). Whatever we each desire, in whatever way, is what we call ‘good’. Hobbes claims that, because our desires are very variable, our assessments of value often lead to conflict, in two important ways. First, different people’s desires frequently conflict, so that we each form conflicting judgments about ‘what is conformable or disagreeable to reason in the actions of common life’ (L 15.40). Second, ‘the same man in divers times differs from himself, and one time praiseth (that is, calleth good) what another time he dispraiseth (and calleth evil)’ (L 15.40). All that makes us consistently agree on what to call ‘good’ and ‘evil’, so that we avoid such frequent conflicts, is the intervention of a strong sovereign: one who makes consistent demands on our actions while keeping the peace, so that we all conform in our desires to perform these actions, and so ensure peace, more than we desire anything else.

Whether or not Hobbes’s subjectivist account of value directly influenced Hume’s response to the variability objection, they address very similar concerns. Hume agrees with Hobbes that what is good is, fundamentally, just what pleases. He too wants to explain why we nevertheless frequently agree about what to call ‘good’, in cases where we each feel quite different pleasures and pains. However, Hume asks only why our verbal evaluations of *characters* are less variable than our violent passions towards them, and his answer does not rest on any desires resulting from our fear of a sovereign’s power.

Consider how we might talk about people, if our evaluative language was equally influenced by the many passions that we feel towards them. Hume hints at two answers, which are strongly reminiscent of the two kinds of evaluative disagreement discussed by Hobbes. First, ‘a man, that lies at a distance from us, may, in a little time, become a familiar acquaintance’ (T 3.3.1.15, SBN 581). In any such case, we would increasingly describe the person in more positive terms over time, because the increase in familiarity would cause us to

feel increasing pleasure. Second, 'every particular man has a peculiar position with regard to others' (T 3.3.1.15, SBN 581). Therefore, we would frequently disagree with one another about whether a person's character was 'good' or 'bad', as we would experience different passions towards it. In each kind of case, there would be 'continual *contradictions*', or at least disagreement, in our evaluative language regarding characters, much as Hobbes describes (T 3.3.1.15, SBN 581).

Hume's discussion of these potential 'contradictions' occurs in the same paragraph in which he introduces the notion of a common point of view. We have seen that he seems to discuss variations in our sympathetically derived, non-moral pains and pleasures in this paragraph, which he contrasts with the uniformity of our moral language. This strongly suggests that he intends the common point of view to explain why our verbal evaluations of characters follow our sentiments of approbation and disapprobation in their uniformity, rather than varying as our other passions do.

To take stock: Hume has argued that all questions of value ultimately reduce to questions of pleasure and pain. Any character trait is good just where we feel some pleasing passion towards it. Hume has also argued, entirely consistently, that all judgements of moral goodness are pleasing sentiments, directed towards character traits. Yet, as Hobbes would surely urge, he must also explain why we call all similarly benevolent motives 'good', despite the fact that many of our non-moral sentiments towards benevolent characters vary from case to case. A benevolent motive may sometimes cause painful passions such as jealousy, anger and so forth, but we do not call it 'bad' in such cases. Why do we ignore *these* sentiments in our 'general decisions' concerning the value of characters, in favour of our more uniform sentiments? (T 3.3.1.16, SBN 582). This is the uniformity problem.

*6.1.2. Hume's response to the uniformity problem*

In §6.2, I will argue that, in his moral *Enquiry*, Hume responds to the uniformity problem by arguing that we find our uniform sentiments so pleasing and useful that we developed moral language to express them alone. In the *Treatise*, however, he develops a significantly more complex response. This is very possibly because he feels compelled by his *Treatise* theory of value to allow that *all* painful or pleasurable passions are, fundamentally, as evaluative as one another. This theory of value entails that for an object to cause a negative feeling just is for that object to be bad. Hume needs to reconcile this with his theory of moral judgement, by which we call characters 'bad' only where we feel *moral* disapprobation towards them. He does so by arguing that we *mistake* our moral sentiments for more appropriately evaluative responses than any others, so that we feel compelled to correct the 'general principle of our blame or praise' – the sum total of our feelings towards characters – by our uniform, calm 'principles', or moral sentiments (T 3.3.1.18, SBN 583).

Hume claims that, whenever we contemplate any distant character towards whom we experience approbation, we feel that we *ought* to experience equally strong, non-moral, 'affection and admiration' towards them (T 3.3.1.16, SBN 582). However, it is 'seldom men heartily love what lies at a distance from them, and what no way redounds to their particular benefit' (T 3.3.1.18, SBN 583). Therefore, Hume argues, we feel obliged to talk *as if* we felt the violent sentiments that we would feel, if only we were nearer. He claims that this desire to use 'the terms expressive of our liking or dislike, in the same manner, as if we remain'd in one point of view', ultimately causes us to express only our uniform moral sentiments (T 3.3.1.16, SBN 582). By expressing only these sentiments, we can 'fix on some *steady* and *general* points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation' (T 3.3.1.15, SBN 581-2).

Hume begins the paragraph immediately following T 3.3.1.15 with the following claim: ‘In general, all sentiments of blame or praise are variable, according to our situation of nearness or remoteness, with regard to the person blam'd or prais'd, and according to the present disposition of our mind’ (T 3.3.1.16, SBN 582). As a ‘general’ claim, this allows that the moral sentiments are exceptions. If we think of a historical character, like Brutus, whose motives are of useful or agreeable types, then we will immediately experience only one strong sentiment: the calm sentiment of approbation. We experience precisely the same degree of approbation towards our similarly motivated servant, but here we also feel violent, non-moral pleasures of ‘love and kindness’ (T 3.3.1.16, SBN 582). Recognising this disparity, we feel obliged to try and summon up similar feelings of love and kindness towards Brutus. However, we are unlikely to achieve this, so we merely talk as if we have done so: we call the two characters equally ‘good’. We do this by *ignoring* all our violent passions, and by expressing only the uniform calm sentiments that motivated us to correct our judgment in the first place. ‘Experience soon teaches us this method of correcting our [violent] sentiments, or at least, of correcting our language, where the sentiments are more stubborn and inalterable’ (T 3.3.1.16, SBN 582).

Although Hume does not say so here, his arguments concerning the causes of moral judgements suggest that our best method of ignoring our violent sentiments is to turn our view *from* the ideas of the particular traits of each person and *to* the general notions of these traits, so as to consider them merely as tokens of their general types. By thinking of any benevolent character *just as* a benevolent character, we can ‘correct the momentary appearances of things, and overlook our present situation’ (T 3.3.1.16, SBN 582).

Hume compares this kind of correction to a case where we call someone ‘beautiful’ when we see him at a distance, because we associate our idea of the distant person with an idea of the beauty that we would see if we were nearer to him (T 3.3.1.15, SBN 582). In both

moral and aesthetic cases, we correct our judgements according to 'sentiments of beauty' that arise via custom (T 3.3.1.20, SBN 584-5, see also T 3.3.1.8, SBN 576-7). Here, our idea of the person's beauty produces a calm aesthetic pleasure, which we express because we feel that only this sentiment responds appropriately to the person's features. Hume claims that this correction occurs via a process of 'reflection' (T 3.3.1.15, SBN 582). I think we should take him to be referring to the introspective nature of this process, such that the correction is caused purely via associations of ideas and 'impressions of reflection', rather than to any process of reflective reasoning (T 2.1.1.1, SBN 275). It is simply not plausible to think that we must always reflect on whether distant people's faces are such as to please when they are closer before we can sincerely call them 'beautiful'. Presumably, therefore, *this* correction occurs purely habitually, and so we should assume the same for the moral case.[6]

Árdal (1966, 120) is surely correct to read Hume as approving of our habit of treating moral value as 'objective', because this habit 'tends to eliminate the friction which arises in our arguments about the value of qualities of character'. Hume claims that, if we did not correct our evaluations, and so avoid continual 'contradictions' (which he must realise are not *really* contradictions), then we could not 'ever make use of language' (T 3.3.1.16, SBN 582). By focusing on only our moral sentiments, we all consider character traits without reference to our own interests, because these sentiments are affected only by the ways that such traits generally influence 'those who have an intercourse with any person' who possesses them (T 3.3.1.17, SBN 582). Hume certainly praises the practice of moralising, and sees it as beneficial (T 3.3.3.2, SBN 603). Nevertheless, in the *Treatise*, he argues that our desire to

[6] Davie (1998, 286-87) gives a similar argument concerning Hume's analogy between the adoption of a common point of view and the visual corrections by which, Hume claims, we come to understand the true sizes of distant objects. People and animals appear to make such corrections without conscious reflection.

correct our evaluations of characters according to our moral sentiments rests on a profound mistake.

Hume claims that such corrections are 'common with regard to all the senses' (T 3.3.1.16, SBN 582). Looking out across a forest, for example, the more distant trees seem tiny, but we correct for this appearance by saying that all the trees are equally 'tall'. We perform this correction because we believe that the trees really are equally tall. Somewhat similarly, we feel less violent pleasure towards a distant, benevolent person than towards a nearer one, and we correct for this by saying that they are equally 'good'. In the *Treatise*, Hume does not appear to recognise that this might be because we all implicitly agree to use certain, specifically moral terms, like 'virtuous', to express *only* our moral approbation. Instead, he argues that we perform this kind of correction because we mistake our calm sentiments of approbations for evaluative beliefs. Considered merely *as* benevolent characters, the nearby character and the distant character feel as pleasing as one another. The kind of feeling involved is so calm, however, that we mistake the relevant passions for beliefs. We think we believe that the similarly benevolent characters are equally good, and so we call them both 'good', even though we really feel that the nearer character is *better*, overall. Hume is, to this extent, an *error theorist* in the *Treatise*: he believes that we only moralise as we do because we hold certain, erroneous beliefs.

In T 3.3.1.18 (SBN 583), Hume argues that, whenever a moral sentiment is caused via general rules and delicate sympathy, we mistake it for a reasoned belief about the real value of a character. What we take to be a process of reasoning is in fact a 'general calm determination of the passions, founded on some distant view or reflection'. (Recall, from §2.2, Hume's claim at T 2.3.4.1 (SBN 419), that calm pleasures are caused by 'remote' goods). Hume is surely referring to that kind of process, caused by a general notion of a character trait, by which general rules and delicate sympathy produce a moral sentiment.

A ‘determination’ is, for Hume, a psychological process that produces ideas by habitual association (T 1.3.12.7, SBN 133, see also T 1.3.12.9, SBN 134 and T 1.3.14.29, SBN 169). Hume has already argued, at T 2.3.3.8 (SBN 417), that we frequently confuse those psychological processes that produce calm passions with the ‘determinations of reason’. His argument in T 3.3.1.18 is very brief, but I propose we understand it as follows. If we think of any person’s motive of, for example, cruelty, then we habitually form a general notion of it: an idea ‘founded on some distant view or reflection’, such that we think of it just as a cruel motive. A ‘determination’ will then produce an idea of the type of pain that we habitually expect cruelty to cause to ‘those, who have any commerce with the person we consider’ (T 3.3.1.18, SBN 583). This can involve no *particular* idea of anyone who has commerce with the cruel person, because Hume is arguing that we will sympathise with this idea, and to the same extent, whether we know the cruel person well or merely read about her. It can only be a quasi-believed general notion of the relevant kind of pain. We sympathise with it, presumably via delicate sympathy, and feel a strong, ‘calm’ disapprobation. This entire process – from considering the motive to experiencing the sentiment – occurs without conscious reflection. The uniformity of our moral sentiments, their calmness, and the habitual, associative manner of their production all cause us to mistake them for reasoned beliefs about the real value of token motives.

However, we *also* experience various different, violent passions towards different cruel people, depending on the ways that they affect us or our friends. We try to ignore these violent passions because we think that ‘reason requires such an impartial conduct’ (T 3.3.1.18, SBN 583). If we read of someone’s cruel action in a history book, and then hear of someone who recently performed a similarly cruel action next door, we will mistakenly think that we *believe* that both cruel motives are equally bad. We feel that reason requires us to correct our violent passions according to the ‘standard of merit and demerit’ that comes from

thinking of each motive merely as a cruel motive (T 3.3.1.18, SBN 583). We therefore express *only* our moral disapprobation towards each cruel motive: we 'blame equally a bad action, which we read of in history, with one perform'd in our neighbourhood t'other day' (T 3.3.1.18, SBN 584).

In the *Treatise*, a *genuinely* corrected sentiment of 'disapprobation' towards a character would be one in which our variable, violent feelings have been brought into line with our uniform, calm sentiment of moral disapprobation (T 3.3.1.18, SBN 584). Here, following his theory of value rather than his theory of moral judgement, Hume uses the term 'disapprobation' to refer to the sum total of any painful feelings felt towards a character. However, this is something of a play on words, I think. Here, Hume is concluding his argument that we only *mistakenly* come to believe that our moral disapprobation is more appropriately evaluative than any other painful sentiment, and that this is the only reason why we correct our evaluation of any character. We feel that we ought to be as angry towards historical cruelty as we are towards cruelty in our neighbourhood, but we cannot meet this standard. We can only talk *as if* we were equally appalled, and we do this by expressing only our moral disapprobation towards each cruel motive.[7]

In T 3.3.3.2 (SBN 602-3), Hume gives a somewhat similar argument to the one at T 3.3.1.18. We saw, in Chapter 2, that he thinks we only instinctively desire to help our loved ones and children, although he allows that we may also feel compassionate desires to help adult strangers, if we experience strong, non-delicate sympathies with their particular pains (T 2.2.7.2, SBN 369). In short, we typically only want to help those near to us, and so Hume

[7] Similarly, when Hume refers back to this argument in T 3.3.1.21 (SBN 585), he describes our focus on our moral approbation over all other pleasing sentiments towards characters as the process of correcting the 'different sentiments of virtue'. However, two paragraphs later, he uses the phrase 'sentiments of virtue' to refer to only moral approbation (T 3.3.1.23, SBN 586).

claims that we have all learned from experience that people generally only help those in their 'narrow circle' (T 3.3.3.2, SBN 602). He then claims that, wherever we contemplate any benevolent or generous person, the 'natural tendency' of her motive will cause us to sympathise with an idea of the happiness of those with any 'particular connexion' to her (T 3.3.3.2, SBN 602). As at T 3.3.1.18 (SBN 583), this idea cannot be a particular one, because we will sympathise with it whether or not we know anything about those who are connected to the benevolent or generous person. We 'forget our own interest' and instead 'consider the tendency' of the relevant motive to cause happiness or unhappiness to those around its possessor (T 3.3.3.2, SBN 602). We 'neglect' all passions that result from our variable sympathetic responses, because we find that these *are* very variable over time and between people, in favour of our uniform 'calm judgments concerning the characters of men': our moral sentiments (T 3.3.3.2, SBN 603).

By this point, Hume has distinguished the violent 'passions' of the 'heart' from the calm sentiments of 'taste', including moral taste (T 3.3.1.23, SBN 586). Here, he claims that the '*heart* does not always take part with those general notions, or regulate its love and hatred by them' (T 3.3.3.2, SBN 603). The 'general notions' in question are those that provide a 'general inalterable standard, by which we may approve or disapprove of characters and manners' (T 3.3.3.2, SBN 603). They are, I take it, general notions of the character traits under evaluation, such as benevolence and generosity. We try to 'regulate' our passions by, for example, feeling violent love towards all benevolent characters, because when they are considered just as benevolent characters, they seem equally good. This is, of course, because we feel equally strong approbation towards each one. We rarely succeed in summoning up violent love in all cases, so we typically express *just* our moral sentiments towards all such traits, so that we may talk as if we have brought our violent passions into line with them. This way of talking is 'sufficient' for moral 'discourse' (T 3.3.3.2, SBN 603).

The sympathetic process described in the two passages just discussed – T 3.3.1.18 (SBN 584) and T 3.3.3.2 (SBN 602-3) – is generally thought to be the corrective exercise at the heart of the Correction Thesis. By my interpretation, it is instead the habitual and unreflective process by which delicate sympathy responds, via general rules, to general notions of character traits. It is this process which ensures that we all feel the same uniform sentiments towards the same characters. In the *Treatise*, then, Hume does not argue for the Correction Thesis. To adopt the 'common point of view' is to express only our moral sentiments, so that we may correct for variations in our non-moral sentiments (T 3.3.1.30, SBN 591). No imaginative exercise is involved.

However, Hume seems somewhat unpersuaded by his own claim that we feel obliged to passionately love or hate distant characters: he acknowledges that our violent passions do not 'often correspond entirely to the present theory' (T 3.3.1.18, SBN 583). Moreover, he sometimes hints that terms like 'virtue' or 'disapprobation' ought properly to apply to *any* pleasures or pains that we feel towards character traits, as at T 3.3.1.18 (SBN 584) and T 3.3.1.21 (SBN 585). Hume usually reserves these terms for purely moral evaluations. However, he has some reason to be wary of doing so. Throughout his *Treatise*, Hume has been arguing for his deeply controversial, Hobbesian theory of value. He may well worry that, if he argues that evaluative terms like 'virtue' or 'disapprobation' should *only* apply to cases where we express calm, uniform sentiments, then this might undermine his argument that there is no fundamental difference between moral sentiments and violent passions like love and hatred.

I suspect that a very similar worry might explain another tension, discussed in §1.1, between Hume's suggestion, at T 3.3.1.27 (SBN 589-90) that we approve of some character traits simply because we find them immediately agreeable, and his considered view, at T 3.3.1.29 (SBN 590), that all *moral* approbation of immediately agreeable traits is caused via

the same kind of sympathetic process that causes our moral approbation of useful traits. Hume seems torn between two 'systems of morality', both of which, he thinks, 'merit our attention': 'sentiments may arise either from the mere species or appearance of characters and passions, or from reflections on their tendency to the happiness of mankind, and of particular persons' (T 3.3.1.27, SBN 589).

Hume freely allows that pleasing sentiments arise in both ways. We often feel passions like joy and love from the 'mere… appearance' of immediately agreeable traits like wit or politeness. We *also* experience approbation wherever we associate that kind of trait with benefitting society or individuals. Hume's question is whether our violent and variable sentiments, like love, should be considered *within* a 'system of morality', such that we count them as moral sentiments. His theory of value suggests that all such passions should be included. Were Hume to accept this, then his moral system would include our love of wit as a moral sentiment, no less than our approval of wit. However, this would undermine his carefully observed theory that only our calm, uniform sentiments form the basis of our most socially important evaluations of characters.

It may be a sign of this tension that Hume's immediate answer to the question is to claim that 'both these causes are intermix'd in our judgments of morals; after the same manner as they are in our decisions concerning most kinds of external beauty' (T 3.3.1.27, SBN 590). He opines that 'reflections on the tendencies of actions have by far the greatest influence, and determine all the great lines of our duty' (T 3.3.1.27, SBN 590). However, he concedes the existence of 'cases of less moment, wherein this immediate taste or sentiment produces our approbation' (T 3.3.1.27, SBN 590). Here, much as at T 3.3.1.18 (SBN 584) and T 3.3.1.21 (SBN 585), Hume uses the term 'approbation' to refer to a non-moral sentiment: here, 'love and esteem' (where, confusingly, 'esteem' is used, as in Book 2, in its

non-moral sense). I take his point to be that we often praise people when we love them for their wit, no less than when we morally approve of them.

Hume ultimately claims that there *is* an important and fundamental distinction to be made between moral approbation and other pleasing passions:

> [H]owever directly the distinction of vice and virtue may seem to flow from the immediate pleasure or uneasiness, which particular qualities cause to ourselves or others; 'tis easy to observe, that it has also a considerable dependence on the principle of *sympathy* so often insisted on (T 3.3.1.29, SBN 590).

Hume believes that we can only explain our uniform approbation towards immediately agreeable virtues if we 'have recourse to the foregoing principles' of, presumably, general rules and delicate sympathy (T 3.3.1.29, SBN 590). He concludes that our delicate, or 'constant and universal', sympathetic pleasures and pains 'are alone admitted in speculation as the standard of virtue and morality. They alone produce that particular feeling or sentiment, on which moral distinctions depend' (T 3.3.1.30, SBN 591). However, in the *Treatise*, he seems uneasy about fully committing to this moral system.

In §6.2, I will argue that, in the *Enquiry*, Hume modifies his account of the common point of view. He no longer argues that we feel obliged to alter our violent sentiments, or that there is any mistake involved in our coming to moralise. He pays greater attention to the social benefits of coming to 'converse together on… reasonable terms' about 'characters and persons' (T 3.3.1.15, SBN 581). This allows him to resolve the tension between his theories of value and of moral judgement, at least to his own satisfaction. He argues that we possess a

unique kind of moral *language*, but that our moral *judgements* are nevertheless passions which are, fundamentally, no different in kind from any other.

### 6.2. The common point of view in the moral *Enquiry*

In the *Enquiry*, as in the *Treatise*, to adopt the common point of view is to evaluate characters purely by expressing our moral sentiments. However, Hume now clarifies and expands on his treatment of moral language. He argues that we express our 'general' sentiments of approbation or disapprobation by using 'a peculiar set of terms' that we have developed *only* for this purpose: moral terms (M 9.8, SBN 274).

Hume argues that we can usually come to agree in our evaluations of characters because we all possess a 'general principle of moral blame and approbation' (M 5.46, SBN 231-2). This 'principle' is our disposition to experience moral sentiments, via 'humanity', or delicate sympathy. It is a 'general' principle in that it produces uniform sentiments. All other types of sentiment are stronger and more violent when directed towards token characters that affect us or our friends than when directed towards ones that affect strangers, so that we feel greater pains and pleasures in the former kinds of case than in the latter kinds. However, we each learn to 'correct these inequalities by reflection, and retain a general standard of vice and virtue, founded chiefly on a general usefulness' (M 5.42n. 25.1, SBN 229).

Hume's language of 'correction', here and elsewhere, is similar to that of the *Treatise*, but not identical. The similarities are perhaps most noticeable at M 5.41 (SBN 227-8). Hume observes that, when we compare a nearby, benevolent statesman to a similarly benevolent one in a distant country, we 'own the merit to be equally great' in both cases, despite feeling a 'more passionate regard' towards our countryman (M 5.41, SBN 227). This is the kind of case that led to the variability objection in the *Treatise*.

As in the *Treatise*, Hume's language in his discussion of this case is by no means as clear as we might wish. Here, this is presumably because he is officially neutral as to whether moral judgements are beliefs or sentiments. However, assuming that to 'own' someone's merit to be great is to express one's moral approval of them, and that to feel a 'passionate regard' is to feel violent, non-moral sentiments, his claim is consistent with Generality. Hume then argues that, because of this disparity, we correct the 'inequalities of our internal emotions and perceptions' (M 5.41, SBN 227). Again, the phrase is ambiguous. Hume could mean that we correct for the initially different degrees of approbation that we experience in each case: in other words, he could argue for the Correction Thesis. Alternatively, he could mean that we judge both statesmen only according to our moral approbation for them, so that we ignore our non-moral 'passionate regard' for our countryman. Only the second of these two claims is consistent with Generality.

In a footnote, Hume claims that such corrections are performed 'by an easy and necessary effort of thought', for the same kind of reason that 'the tendencies of actions and characters, not their real accidental consequences, are alone regarded in our moral determinations or general judgments' (M 5.41n. 24.1, SBN 228). Just as in cases of virtue in rags, we offer the same 'general praise' to any action or character of a type that tends to cause happiness, regardless of our 'real feeling or sentiment'. This certainly suggests that we evaluate characters by expressing only our uniform moral sentiments, while ignoring our violent, non-moral sentiments.

As in the *Treatise*, Hume claims that we frequently correct our visual perceptions by reference to beliefs about the real sizes of objects, and he argues that this is roughly analogous to the corrections that we make when we verbally evaluate motives. However, unlike in T 3.3.1.18 (SBN 583-4), he does not argue that we correct these evaluations because we mistake our calm sentiments for evaluative beliefs. We do so just because we want our

verbal evaluations to be uniform. Both the visual and the evaluative corrections occur where a variation of some kind impedes our desire to 'think' and 'talk steadily' (M 5.41, SBN 228). The possibility of error does not seem relevant to the comparison that Hume wishes to draw: he only talks of 'error' in the case of visual perception (M 5.41, SBN 227). The point of the analogy is to show how 'fluctuating situations produce a continual variation on objects, and throw them into such different and contrary lights and positions' that we find it difficult to think or talk about them in any consistent manner (M 5.41, SBN 228).

Hume claims, as at T 3.3.1.15 (SBN 581), that our evaluations of characters would vary, between people and over time, were we to evaluate them in accordance with the many 'sentiments… which have a reference to private good' (M 5.42, SBN 228-9). Any population that relied on *all* their passions when they judged one another's characters would rarely agree in their judgements. Indeed, because of their very different interests and feelings, their conversations about such matters would be barely 'intelligible' (M 5.42, SBN 228).

Hume argues that we have avoided this problem by relying only on our 'general preferences and distinctions' (M 5.42, SBN 228). We realise that we all consistently prefer some types of character traits over others, in a calm but uniform way. According to Hume, of course, this is because we have sentiments that consistently 'attach the notion of good to a beneficent conduct, and of evil to the contrary' (M 5.42, SBN 229). We come to rely on these sentiments whenever we judge character traits, because doing so allows us to agree in our approval of generally beneficial traits and our disapproval of generally harmful ones. This is both pleasing and useful, so that we learn to ignore our variable, violent passions in our evaluations of character. Unlike in the *Treatise*, Hume's argument does not rest on a claim that we feel obliged to *alter* our violent passions. In a modified version of his discussion at T 3.3.3.2 (SBN 602-3), he still acknowledges that 'the universal, abstract differences of vice and virtue' rarely cause us to change the violent passions of the 'heart' (M 5.42, SBN 229).

However, he no longer sees this as a problem for us, to be overcome by talking *as if* they have done so. He simply argues that the moral passions fully meet our social requirements when we publicly evaluate characters.

Unlike in the *Treatise*, then, Hume does not claim that we want to correct our evaluations of character *because* we confuse moral sentiments for evaluative beliefs. It is, however, very possible that he still thinks that the similarities between the above two kinds of corrective processes cause us to mistake our moral sentiments for evaluative beliefs. Presumably such mistakes are made even more likely by their frequently occurring after we have used reason to 'instruct us in the tendency of qualities and actions, and point out their beneficial consequences to society and to their possessor' (M App. 1.2, SBN 285). Hume is clear that all such reasoning serves only to 'pave the way' for moral sentiments, by causing beliefs about the tendencies of character and action types to produce pleasure or pain, for example (M 1.9, SBN 173). Reasoning may allow us to correctly categorise a token motive, or it may, presumably over some time, cause us to change some of our habitual associations between certain motive types and relevant effects. However, no reasoning is *required* for us to form consistent and socially useful moral judgements. The only necessary processes for this are those of our habitual associations of ideas, delicate sympathy, and moral sentiment.

In fact, Hume makes this point more clearly in his first *Enquiry* than he does in his second:

> Morals and criticism are not so properly objects of the understanding as of taste and sentiment. Beauty, whether moral or natural, is felt, more properly than perceived. Or if we reason concerning it, and endeavour to fix its standard, we regard a new fact, to wit, the general taste of mankind, or some

> such fact, which may be the object of reasoning and enquiry (E 12.33, SBN 165).

Despite placing a greater emphasis on morally relevant forms of reasoning in the moral *Enquiry* than in the *Treatise*, Hume still assumes that people within any one community will only rarely come to different moral judgements about the same kinds of character traits. Indeed, he sometimes assumes that we *all* associate the same kinds of traits with the same kinds of pleasures or pains, so that the moral sentiments operate entirely consistently between people: 'Whatever conduct gains my approbation, by touching my humanity, procures also the applause of all mankind, by affecting the same principle in them' (M 9.8, SBN 274).

Whether plausible or not, this assumption is very useful for Hume, because it allows that approbation, along with disapprobation, can help us agree in our 'calm judgments and discourse concerning the characters of men' (M 5.42, SBN 229). If we all focus only on our calm, uniform sentiments, then we all approve of motives that tend to be socially beneficial and disapprove of motives that tend to be socially disruptive. As suggested in Chapter 5, Hume argues that this is so useful and pleasing to us that we have developed a kind of language to express only our moral sentiments:

> General language… being formed for general use, must be moulded on some more general views [than those of our private interests], and must affix the epithets of praise or blame, in conformity to sentiments, which arise from the general interests of the community (M 5.42, SBN 228).

All and only those sentiments that are produced via humanity arise from 'general interests', in that they are experienced by all members of the community, uniformly, towards all character traits of generally pleasing or displeasing kinds. As in the *Treatise*, Hume allows that there is no fundamental distinction to be made between our moral and non-moral passions. What *has* changed is that he has increased his focus on moral language. This allows him to fully reconcile his claim that moral judgements are fundamentally like all other pleasures or pains with his arguments that they play a central role, unlike that of any other passion kind, in our evaluations of character. Although the moral sentiments are of the same general kinds as those passions caused by the 'real accidental consequences' of actions and characters, the language in which we express them is of a very different kind (M 5.41n. 24.1, SBN 228).

For example, I may be pleased by any action that satisfies my 'vanity' or 'ambition', but this pleasure will 'not [have] a proper direction' for me to treat it as a *moral* evaluation (M 9.5, SBN 271). It is merely a pleasure at my own particular benefit, and those with whom I am conversing are unlikely to feel similarly pleased, or to love the actor as I do, unless they happen to be similarly benefitted. Nevertheless, the same action may also cause approbation, via humanity. I *will* expect my interlocuter to share this sentiment, and I will give it the status of a moral judgement:

> [The] affection of humanity may not generally be esteemed so strong as vanity or ambition, yet, being common to all men, it can alone be the foundation of morals, or of any general system of blame or praise. One man's ambition is not another's ambition; nor will the same event or object satisfy both: But the humanity of one man is the humanity of

> every one; and the same object touches this passion in all human creatures (M 9.6, SBN 273).

Of course, we may still express violent, non-uniform sentiments when we evaluate people's characters, but not via moral language:

> When a man denominates another his *enemy*, his *rival*, his *antagonist*, his *adversary*, he is understood to speak the language of self-love, and to express sentiments, peculiar to himself, and arising from his particular circumstances and situation. But when he bestows on any man the epithets of *vicious* or *odious* or *depraved*, he then speaks another language, and expresses sentiments, in which, he expects, all his audience are to concur with him (M 9.6, SBN 272).

We choose whether or not to evaluate people in moral terms. Hume claims that anyone who chooses to use moral language to express only their uniform sentiments has, in virtue of making this choice, 'chosen [the] common point of view, and… touched the principle of humanity, in which every man, in some degree, concurs' (M 9.6, SBN 272). Again, nothing like the Correction Thesis is involved in Hume's discussion of this common point of view.

Here, I conclude my examination of Hume's theory of moral sentiments. In Part 2, I will ask what we in the 21st century can learn from it.

# Introduction to Part Two

> The hypothesis which we embrace is plain. It maintains that morality is determined by sentiment. It defines virtue to be *whatever mental action or quality gives to a spectator the pleasing sentiment of approbation;* and vice the contrary. We then proceed to examine a plain matter of fact, to wit, what actions have this influence: We consider all the circumstances, in which these actions agree: And thence endeavour to extract some general observations with regard to these sentiments. (M App1.10, SBN 289).

After much consideration of Hume's metaethics, I now want to apply some of his theses to the field of contemporary metaethics. This is not a new ambition, of course; there are many 'Humean' theories and views to be found in 21st century moral philosophy. Indeed, there are several different ways that such Humean views may be formed, and so more than one way that we could try to learn from Hume.

One way would be to directly translate as much of Hume's moral sentimentalism as possible into acceptable terms for the 21st century, so that it generally complies with contemporary theoretical constraints. This might lead to a theory, perhaps rather like Slote's (2010), in which moral approval is a pleasant feeling caused by our empathy with others. Alternatively, we could consider Hume's wider philosophical commitments and arguments, and then aim to develop a metaethical view that relies on these insights at the expense of the details of his sentimentalist theory. This is the approach of Blackburn (1993a, 167), who once described his expressivist theory as 'a modern version of Hume's theory of the nature of ethics, but without any commitment to particular operations of passions such as sympathy.' As a third possibility, we could look to Hume the psychologist over Hume the philosopher,

and ask how he might inform our understanding of the psychological processes involved in moral judgement, according to our best, empirically informed, understanding of these. This is the approach of Haidt (2012, 116), who once claimed that Hume 'laid a superb foundation for "modern science," one that has, in my view, been largely vindicated by modern research'.

My approach will involve elements of all three of the methods listed above. However, I intend to set myself achievable ambitions, and so I will be constrained in at least two ways. First, I cannot begin a theory entirely anew. I therefore aim to work with, and to build on, some recent and well-developed theories, particularly those developed by Haidt (2012) and Gendler (2008a; 2008b). Following Haidt, I endorse both 'moral intuitionism' and 'Moral Foundations Theory', although the latter will be of lesser importance to my theory. Here, I do not intend 'moral intuitionism' to refer to theories, like Ross's (1930), by which the truth of some core set of moral propositions is self-evident. Instead, I mean this term to refer to the thesis that moral judgements are typically or always produced by intuitive rather than by reflective thought processes. I will explain this in detail in Chapter 7. Following Gendler, I will endorse the theory of 'alief'. I will explain Haidt's theories in chapter 7, and Gendler's in chapter 8. I will add to Haidt's arguments for moral intuitionism in chapter 8.

My second constraint is due to the fact that I cannot hope to consider the viability of every aspect of Hume's theories. I will therefore focus on his thesis of Generality, as discussed in §3.1.2, and on the automatic, associative processes that he claims are involved in the causes of moral judgements.

I have chosen these two constraints because they are closely related. Haidt and Gendler, along with many others, are involved a recent resurgence in attempting to understand our automatic, associative thought processes. Indeed, as Gawronski and Bodenhausen (2006) claim, a 'major theme' of current psychological research is the notion of automaticity:

> Many aspects of human behavior that have previously been assumed to have their roots in higher order processes of deliberate reasoning are now viewed as resulting from automatic processes that may occur spontaneously and outside of people's awareness or control. (Gawronski and Bodenhausen 2006, 692)

I will argue, as Haidt does, that moral judgement is one such aspect of human behaviour. I will use Gendler's theory of alief as the framework within which to argue for this. In so doing, I will follow Sinnott-Armstrong and Wheatley's (2014, 452) deliberately broad definition of 'moral judgement': 'A judgment is a moral judgment if and only if the judger does, or would, think of it as similar enough in relevant respects to judgments that that judger takes to be exemplars or paradigm cases of moral judgments'.[1] If someone calls an action 'wrong', and they do so because they take it to be wrong in much the same way that they take

[1] I think that, for any theory of the psychological nature and causes of moral judgements, a broad definition of 'moral judgement' is required, to avoid excluding candidates merely by terminological fiat. Haidt (2001, 817) similarly employs a broad definition, defining moral judgements as 'evaluations (good vs. bad) of the actions or character of a person that are made with respect to a set of virtues held to be obligatory by a culture or subculture'. Here, it seems that 'virtue' is used purely descriptively, just to mean some kind of action that is treated as obligatory for all people, or for some class of people, within a certain culture or subculture. I think Sinnott-Armstrong and Wheatley's definition is preferable to Haidt's, so I will endorse this throughout. However, so far as I can see, nothing of importance rests on which of these two definitions I employ.

paradigmatically morally wrong actions to be wrong, then the mental state that plays the relevant causal role in their utterance of this statement is a moral judgement.[2]

In what follows I will, broadly, address two questions: How should we understand moral judgements themselves? And how should we understand the meanings and pragmatics of moral terms and language? In the spirit of Hume, I will assume throughout that we require naturalistic answers to all such metaethical questions: answers that only invoke properties and objects of kinds that are acceptable within scientific explanations. However, unless stated otherwise, I will reject Hume's terminology in favour of contemporary terminology. I will henceforth focus more on our judgements of actions than of character traits, although I will consider both.

In the first two chapters of Part 2, I consider the psychological causes of moral judgements and, as Hume suggested in the quotation with which I began, 'endeavour to extract some general observations with regard to these sentiments': notably, that we approve of token actions and character traits of types that we associate with causing happiness (broadly construed) and that we disapprove of token actions and motives of types that we associate with causing harm (broadly construed). In the final two chapters, I at least begin developing a theory of moral language, by which the meaning of a moral term like 'virtue' may indeed be defined as '*whatever mental action or quality gives to a spectator the pleasing sentiment of approbation*'. Like Hume, however, I argue that such simple definitions bely the fact that moral language plays an important and complex pragmatic role in our social lives.

[2] It is very hard to define the 'relevant' causal role here. I assume that, whatever the relation is between moral judgements and sincere moral utterances, there is some relevant causal role between the former and the latter, such that the former plays a crucial causal role in the occurrence of the latter.

In Chapter 7, I provide some background and context to my arguments, along with an overview of Moral Foundations Theory and moral intuitionism. In Chapter 8, I argue that, given our current state of knowledge, we should understand paradigmatic moral judgments as 'occurrent aliefs': mental occurrences with representational, affective and behavioural components, caused via processes of which the aliever is not consciously aware (Gendler 2008a).

Then, in the final two chapters, I defend an 'emotive subjectivist' account of moral language, which marries a simple subjectivist theory of meaning with a more complex, 'emotivist' account of the pragmatics of moral language. In Chapter 9, I set out the details of theory. I also argue that we can plausibly endorse an 'opacity thesis', by which much of our moral thinking is unavailable to conscious reflection, and that this allows for ambiguities in the meanings of moral terms that we pragmatically exploit in moral discussion. In Chapter 10, I conclude that emotive subjectivism has the resources to vindicate ordinary moral language and practice by borrowing strategies from Blackburn's 'quasi-realist' expressivism, although I also argue that we must rethink some our assumptions about what ordinary moral language and practice looks like.

# 7. Moral Intuitions and Moral Foundations Theory

> It may now be ask'd *in general*, concerning this pain or pleasure, that distinguishes moral good and evil, *From what principles is it derived, and whence does it arise in the human mind?* To this I reply… that 'tis absurd to imagine, that in every particular instance, these sentiments are produc'd by an *original* quality and *primary* constitution... Such a method of proceeding is not conformable to the usual maxims, by which nature is conducted, where a few principles produce all that variety we observe in the universe, and every thing is carry'd on in the easiest and most simple manner. 'Tis necessary, therefore, to abridge these primary impulses, and find some more general principles, upon which all our notions of morals are founded (T 3.1.2.6, SBN 473).

In this chapter, I will address a potential worry for my theory in advance. In Chapter 8, I will argue that any paradigmatic moral judgement is a certain kind of occurrent alief. Moral Foundations Theory (MFT) states that there are several distinct moral 'foundations', so that different moral judgements may take different kinds of action or character trait as their objects (Graham et al. 2009; Graham et al. 2011; Haidt 2012; Haidt and Graham 2007). This may suggest that, as Sinnott-Armstrong and Wheatley (2014) argue, moral judgement cannot be plausibly understood as a single kind of mental state or occurrence. I will argue, against this claim, that MFT is entirely compatible with the view that moral judgements are of a single, unified kind, even where they stem from different foundational areas of concern.

My conclusion in this chapter will be limited to the claim that typical moral judgements plausibly have a common, unifying property. I will argue for this by applying a

fairly general thesis of Humean moral learning to MFT's understanding of moral judgement. MFT at least appears to suggest that different moral judgements are produced by different psychological systems, that we acquired at different points in our evolutionary past. I will argue, as Hume does in the passage quoted above, that it is more plausible that moral judgements have the unifying property of being *learned* in some particular way than that different moral judgements are produced in distinct ways by different kinds of 'original instincts' (T 3.1.2.6, SBN 473).

We saw, in Chapters 3, and 6, that Hume argues that different types of traits may cause us non-moral pleasures in very different ways, by being either useful in some way or agreeable in some way, but that all traits that we associate with causing non-moral pleasures in any way will cause *moral* approbation via only one kind of habitual, unreflective process: that of general rules and delicate sympathy. I will argue for something broadly similar.

I accept MFT's claim that there are several distinct, foundational psychological systems, or 'foundations', which may dispose us to treat very different types of action (or character trait) as morally right or wrong. However, I will argue that this is compatible with the claim that there is only one kind of moral learning. MFT allows that we may learn to respond morally to any type of action that is appropriately related to any (one or more) of the foundational areas of concern. Nevertheless, this allows that all and only paradigmatic moral judgements may share the property of being produced by at least broadly the same kind of process. Indeed, I will argue for an account of just such a process: an intuitive process which occurs, in response to any token action of any type that is appropriately related to any of the foundational areas of concern, *if and only if* actions of that type are associated, by the judger, with causing (some relevant kind of) happiness or unhappiness to those around her.

I will argue for this conclusion mostly independently from my arguments, in Chapter 8, for a 'moral alief theory'. This is so that each argument may stand or fall on its own merits.

In particular, if we wish to reject the moral alief theory, I believe that we may still apply some broadly Humean theory of moral learning to MFT's understanding of moral judgement, and so allow for a unifying property for all typical moral judgements.

In §7.1, I provide an overview of MFT, focusing only on those aspects of the theory that will be relevant to my arguments to come. In §7.2, I provide a similar overview of what I call 'moral intuitionism': the thesis that moral judgements are typically or always produced by intuitive rather than by reflective thought processes. This thesis is generally accepted by the proponents of MFT. In §7.3, I argue that typical moral judgements plausibly have a unifying feature, in virtue of the habitual way in which we learn to respond morally to certain action and character kinds. Therefore, we cannot infer, from MFT, that paradigmatic moral judgements may not form a unified psychological kind. This will clear a path for me to argue, in Chapter 8, that all paradigmatic moral judgements are occurrent aliefs of a unified psychological kind. Finally for this chapter, in §7.4, I suggest that, if typical moral judgements share a unifying feature in the way that I suggest, then this can explain an otherwise puzzling feature of Haidt's research into intuitive moral judgements: our tendency to claim that actions we deem morally wrong are harmful, even when they clearly are not.

### 7.1. Moral Foundations Theory

MFT is a theory about the psychological systems that ultimately cause human beings to make the moral judgements they do (Graham et al. 2009; Graham et al. 2011; Haidt 2012; Haidt and Graham 2007). Clearly, our education and cultural influences play a large part in determining what types of actions (or character traits) we evaluate in moral terms. According to MFT, there are several psychological systems, or 'foundations', that render action types morally salient, although we are not all equally influenced by each of these foundations.

As Haidt (2012, 181) summarises it, MFT 'says that there are (at least) six psychological systems that comprise the universal foundations of the world's many moral matrices'. Each of these systems involves 'innate but modifiable mechanisms'; each has a unique evolutionary history; and each plays a fundamental causal role in the development of some set of locally recognised virtues and vices (Graham et al. 2009, 1030). For example, the Care/harm foundation is that set of psychological mechanisms, largely involved in empathy and parent-child attachment, that ultimately produces the 'widespread human concern about caring, nurturing, and protecting vulnerable individuals from harm' (Graham et al. 2009, 1031). It is thus the foundation of all those moral judgements that concern harmful actions or motives.

So far as I am aware, no MFT theorist has claimed that the psychological mechanisms that are involved in empathy, attachment and so on are directly implicated in the causes of each individual judgement that a harmful action is morally wrong. MFT certainly does not claim that *only* these mechanisms are involved in such judgements (Haidt et al. 2015). Instead, the thesis is the more general one that humans only judge harmful actions to be wrong because we have evolved the relevant foundational set of psychological mechanisms. These mechanisms cause people to be distressed by the harms of others in a way that has led, over time, to their taking actions that harm others to be salient for moralising and liable to be disapproved of.

Other foundations include Liberty/oppression, Fairness/cheating, Loyalty/betrayal, Authority/subversion and Sanctity/degradation (Haidt 2012, 153-154). Without giving detail to them all, some further examples will help. The Sanctity/degradation foundation is or involves the psychological systems that produce disgust, and it leads to the types of moral framework in which it is wrong to enter a temple without the right clothes, or to behave in sexually unconventional ways (Haidt 2012, 103-4; 148-150). Again, for this area of moral

concern, it is only because people have evolved the psychological mechanisms of disgust that they take such actions as salient for moralising, so that they are likely to morally disapprove of such actions.

Authority/subversion is closely related to the kinds of instincts that move dogs and chimpanzees to find and maintain a place in an ordered hierarchy. It is the foundation of moral disapproval towards any threat to one's social order, such as rebellion or disrespect to authorities (Haidt 2012, 142-144). Had human beings never evolved such instincts, according to MFT, nobody would ever disapprove of rebellion or disrespect to authorities as such.

These foundational areas of concern are understood to be unevenly distributed across moral judgements, in two ways. First, for any individual, some foundations may be more prominent than others. Concerns about Care/harm are typically more morally salient to us than other foundational concerns (Haidt et al. 2015). Second, liberals, and only liberals, typically endorse only virtues that result from the Care/harm, Liberty/oppression and Fairness/cheating foundations (Haidt 2012, 184).

A 'liberal' is here understood as a typically highly educated person in a democratic, Western society, with at least a relatively liberal political outlook. Liberals are very unlikely to think of an action as morally wrong unless they consider that action to be potentially or actually harmful, oppressive or unfair. However, this appears to be very unusual behaviour compared to most people in the world, who will typically morally disapprove of at least some actions just because they deem them disgusting, disrespectful or disloyal.

I am a liberal myself, as are most people who read or write academic philosophy or psychology in the West, according to Haidt (2012, 96-97). Moral judgements appear to many of us to share the feature of being somehow concerned with actions that cause, prevent or allow harm. However, psychological research indicates that this appearance is deceptive. Empirical evidence demonstrates that many people judge some actions as wrong, such as an

act of consensual incest or of the desecration of a national flag, even where they explicitly allow them to be harmless (Graham et al., 2009; Graham et al., 2011; Haidt, 2001; Haidt, 2012; Haidt & Graham, 2007). According to MFT, most liberals are very unlikely to consider an action wrong just because it is disgusting, disrespectful or disloyal (in the sense that being unpatriotic may be conceived as 'disloyal'). Non-liberals, however, are very likely to consider some actions wrong for just these kinds of reasons.

MFT is clearly not a theory that stresses the commonalities between moralisers or between moral judgements. According to MFT, moral judgements are not unified by any one kind of foundational concern. The judgement that is wrong to smack children is only related to concerns about harm, whereas the judgement that it is wrong to privately use one's country's flag to clean a toilet is unrelated to such concerns. Moreover, these two judgements are, in some sense, founded on different psychological systems. The former is related to Care/harm; the latter is not. Although liberals and non-liberals presumably share the psychological system that causes non-liberals to disapprove of using a national flag to clean a toilet (as I will discuss further in §7.3), this system only makes such actions morally salient to non-liberals. It appears to play little or no role in the *moral* psychology of liberals.

It is therefore tempting to see MFT, as Sinnott-Armstrong and Wheatley (2014) do, as providing strong support for the thesis that there are no commonalities between all moral judgements: that moral judgements have *no* unifying feature. Here, I follow Sinnott-Armstrong and Wheatley (2014, 454) in understanding a 'unifying feature' of moral judgements as one that can 'support generalizations that hold for all moral judgments and only for moral judgments'. Any unifying feature must also be distinct from the trivial feature of being categorised by us as moral judgements. Sinnott-Armstrong and Wheatley provide a detailed survey of moral judgements – including their content, phenomenology, associated brain states, force, form, and function – and argue that there are no unifying features of moral

judgements. In other words, after much careful consideration, they see no evidence that there is anything, over and above being classified by us as a 'moral' judgement, that all and only moral judgements have in common.

Several theorists have recently challenged this claim, by arguing, against MFT, that people only sincerely call any token action 'wrong', even one like the private desecration of a national flag, because they take it to be such as to cause or to risk harm. For example, Schein and Gray (2015, 1149) suggest that, wherever someone appears to judge an action wrong because they think it degrading, they do so only because they think that the action in question is degrading in some way such that it will cause or risk harm. More generally, they argue that, given a sufficiently broad understanding of 'harm', people only morally disapprove of actions where they are concerned that the actions are harmful (Gray et al., 2014; Schein & Gray 2015; Schein & Gray 2018).

De Villiers-Botha (2020) similarly suggests that, even where people are unconcerned about interpersonal harm, they might be concerned about harm to other kinds of victims, like the self, gods, or communities. DeScioli et al. (2012, 143) also believe that 'victims' are 'central to moral judgments'. They allow that we may often judge an action wrong *before* we locate a victim, but they argue that we then apply a psychological 'moral model' to the action, which includes understanding it as harming a victim of some kind (DeScioli et al. 2012, 148). In many cases, they suggest, this requires us to 'fabricate' victims (DeScioli et al. 2012, 147).

I cannot disprove these claims here, but I find them unlikely. We have seen that Hume focused on several cases where we judge harmless, victimless actions wrong, and there do seem to be many such cases. Drawing on the research behind MFT, Sinnott-Armstrong and Wheatley (2014, 458) provide a short list of actions that may be judged wrong even if they are known to be harmless, such as disrespecting or disobeying one's parents or elders, for

example, which may be disapproved of even if one's parents or elders are known to be unaware of, or even pleased by, such disobedience. It is fairly easy to think of other examples. Here, I assume that MFT is correct in claiming that not all token actions that are judged wrong are taken, by the judger, to harm victims. Nevertheless, as Haidt et al. (2015) stress, this leaves much to be said about the psychology of individual moral judgements. MFT states that individual moral judgements may address different fundamental types of concern, but it does not state any precise role for the foundational systems in the production of individual moral judgements (Haidt et al., 2015).

Indeed, MFT suggests some important similarities between individual moral judgements. MFT accepts that, regardless of their foundations, moral judgements are typically produced *intuitively*: we typically evaluate actions and character traits 'without any conscious awareness of having gone through steps of search, weighing evidence, or inferring a conclusion' (Haidt & Bjorklund, 2008, 188). This is the thesis I call 'moral intuitionism'. I will consider this thesis in §7.2, before arguing, in §7.3, that moral intuitionism suggests that moral judgements share a unifying feature, in virtue of the psychological mechanisms that lie behind moral learning, and so behind individual moral judgements. This is, I argue, consistent with MFT. It is also a species of unifying feature that Sinnott-Armstrong and Wheatley do not address. Before considering this, however, I must address moral intuitionism more generally.

### 7.2 Moral intuitionism

As I will use the term 'intuition', it is to be contrasted with reflective thinking, rather than with a posteriori reasoning, or reasoning more generally. Reflective thinking is thinking that requires conscious attention. In contrast, intuitive thought processes are automatic or 'autonomous': they do not require controlled attention or conscious thought, because they

automatically produce outputs of some kind whenever the relevant stimuli are encountered (Evans and Stanovich 2013, 236). Intuitive thought processes are typically (and perhaps always) unavailable to conscious awareness, either during their occurrence or afterwards, whereas the products of such processes may or may not be available to conscious awareness (Evans and Stanovich 2013; Gendler 2008b, 557). As Nagel (2014, 224) notes, this use of the term 'intuitive judgement' – to mean a judgement formed immediately and without any reflective reasoning – can be traced back to at least Locke's *Essay Concerning Human Understanding*.[1]

Given this meaning of 'intuition', I use 'moral intuitionism' to refer to the thesis that moral judgements are typically or always produced by intuitive rather than by reflective processes. I take it that Haidt is a moral intuitionist in this sense, although he describes his theory a 'social intuitionist' one, because he allies his thesis of moral intuitions to a claim that moral reasoning typically occurs socially rather than individually (Haidt 2001; Haidt and Bjorklund 2008; Haidt 2012, 48). Haidt (2001, 817) says that the 'central claim of the social intuitionist model is that moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning'. Reflective moral reasoning is typically only 'needed', and so only employed, *after* a judgement has been made, if we want to justify or defend our judgement, as when we are challenged on it (Haidt & Bjorklund, 2008, 189-194).

I will return to the topic of conscious moral reasoning in Chapter 8. For now, I will not address this topic, and I will not use the term 'social intuitionism' because I do not want to commit myself to everything entailed by it. Instead, I follow Haidt merely in being a moral

[1] Hume roughly followed Locke in this language, by describing 'relations [that] are discoverable at first sight' as falling 'under the province of intuition' (T 1.3.1.2, SBN 70). He would not have allowed that moral judgements are produced intuitively, in this sense.

intuitionist. I take the central claim of moral intuitionism to be that moral judgements are typically produced via purely intuitive processes, which may or may not then be reflectively endorsed (Haidt 2012, 44-51). I will aim to explain this now.

There is much debate about the nature of intuitive judgements and processes, and about how these may be distinguished from reflective judgements and processes. I do not intend to enter such debates, and I will generally follow Gendler's and Haidt's understanding of such matters.[2] Here, I will consider the core elements of recent theories of intuitive thought, and then examine how Haidt applies these to moral judgement.

Nagel (2014) cites Sloman (1996), who presents two anagrams which helpfully demonstrate some of the differences between reflective reasoning and intuitive reasoning. If you are asked to solve an easy anagram such as 'involnutray', and if you are a reasonably accomplished reader of English, you will very likely only use intuitive reasoning. It will probably seem to you that, simply by looking at these letters, you think of the word 'involuntary'. The solution to the anagram comes immediately to mind, and the process by which it does so is automatic and unavailable to consciousness. As such, it is an intuitive judgement, as understood here.

When we are asked to solve a harder anagram such as 'uersoippv', however, the case is different:

> [We] go through partially introspectable cycles of mental activity, experimenting with different consciously available contents involving various transpositions of letters until we find the solution. [This] type of

[2] Gendler and Haidt sometimes use different terminologies. However, I do not see that there is anything incompatible in their various theses.

> thinking depends on working memory not just to present the original problem, but also to take us via a series of stages towards a solution. Each of these individual stages is itself intuitive, just as Locke observed. (Nagel 2014, 227)

Nagel (2014, 226) focuses on working memory because it is intimately connected to conscious thinking. The key point here is that, while we do not require or employ any conscious attention to solve an easy anagram like ‘involnutray’, we will probably need to apply some degree of conscious attention to the second anagram, before we can see that it is an anagram of ‘purposive’. Intuitive processes are involved in both kinds of case, although they are typically only noticeable to us where they are not operating as elements of more complex, partially reflective processes.

Much research into intuitive processes focuses on ‘implicit attitudes’: those which ‘manifest as actions or judgments that are under the control of automatically activated evaluation, without the performer's awareness of that causation’ (Greenwald et al. 1998, 1464). These are the kinds of attitudes that may cause us to unconsciously expect someone of a certain gender or skin colour to be incompetent or untrustworthy, even where we reflectively believe that such sexist or racist attitudes are reprehensible (Gendler 2011). Such implicit biases are liable to occur in the minds of all of us who live in societies that frequently depict certain groups in a negative (or positive) light.

As the example of implicit bias may suggest, the automatic and non-conscious aspects of intuitive processes make it difficult for us to alter or to prevent the occurrence of implicit attitudes. In many cases, our behaviour is influenced by implicit attitudes without our being consciously aware of any thought processes during their production or of the attitudes so

produced. We may only be able to say afterwards that something or someone did not feel quite right, without being able to explain why.

Indeed, our motivation is very often influenced by feelings of positive or negative affect that may go unnoticed at a conscious level. Lebrecht et al (2012) call such feelings 'micro-valences'. They argue that, even where one is engaged in the seemingly coldly unemotional task of choosing a mug from a cabinet, one is likely to be motivated to choose one mug over the others just because one experiences the most pleasing micro-valence from it (Lebrecht et al 2012; see also Zajonc and Markus 1982, 128). It is unlikely that many micro-valences are consciously noticed or considered in such cases. As Cunningham et al. (2004) demonstrate, the brain states that correlate with valences are activated considerably more frequently than we might expect, and they may activate regardless of our intentions. This suggests that emotional 'arousal/intensity and valence are basic aspects of evaluative processing that likely occur automatically' (Cunningham et al. 2004, 1723). That is, it appears likely that any kind of evaluation will involve, as a central feature, some kind of valence or affect – some barely perceptible kind of positive or negative feeling, in other words – that will occur automatically.[3]

So understood, intuition and micro-valence are similar in at least two ways: they are at least typically unnoticed by us at a conscious level, and they appear to have the psychological function of facilitating quick reactions or decisions without requiring complex

[3] I do not see that there are any important differences between terms like 'valence', 'affect' or 'feeling' in such contexts. What matters is that they are always either positive or negative, so broadly pleasurable or painful, as Hume suggests. As such, I take it that they must be to some extent available to consciousness. However, we may not be able to consciously identify them as individual pleasures or pains, aside from our general awareness of our overall state of positive or negative feeling.

cognition. Indeed, intuitive processes often cause barely perceptible feelings, as Haidt (2001, 818) stresses.

These are merely brief discussions of very complex kinds of mental states and process, about which there is still much debate and much that is unknown. However, to summarise my understanding of these:

*Intuitive* processes are ones that do not require conscious attention. A typical intuitive process is, *inter alia,* an automatic and associative one, which involves or produces some positive or negative feeling (or valence, or micro-valence, or affect, or ….), which may be very minimal and so unnoticed by conscious reflection.[4]

An *associative* process is, much as Hume describes, one where any occurrence of a particular mental representation (which may be produced by perception or otherwise, and which may or may not be consciously accessible) reliably causes a different mental state or occurrence. Typically, an associative process is due to some prior process of unconscious learning. An example of such an associative process would be that by which the mental representation of a mug causes a pleasing micro-valence to occur, because of some previous pleasant experience(s) involving that mug or one resembling it.

*Automatic* processes are, in this context, associative processes that occur regardless of any conscious effort or volition, as where the mere sight of the mug automatically produces the relevant micro-valence. I assume that automatic processes are typically unavailable to conscious attention, but that the products of such processes may or may not be available to conscious attention.

[4] This understanding is due, in large part, to Gendler's (2008a; 2008b), which she discusses in her arguments for the existence of 'aliefs'. I will discuss her theory in detail in Chapter 8.

Reflective processes require conscious attention, but they typically or always include one or more intuitive processes. However, in what follows, I will generally assume a broad dichotomy, between intuitive (purely automatic and purely associative) processes and reflective (neither purely automatic nor purely associative) processes. I recognise that not all intuitive processes may be automatic or associative, but I take it that they very typically are, so that it should be relatively unproblematic to make this assumption here.

With this context in mind, we can consider Haidt's (2001) distinctions between moral 'reasoning' – by which he means reflective reasoning – and moral intuition:

> The most important distinctions… are that intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, whereas reasoning occurs more slowly, requires some effort, and involves at least some steps that are accessible to consciousness. (Haidt 2001, 818)

To say that the outcomes of moral intuitions are 'accessible to consciousness' is not to say that they are all consciously noticed, or at least not to the extent that we are likely to report their occurrence. Haidt (2012, 45) suggests that, throughout even the most mundane of activities, we may experience 'many tiny flashes of condemnation that flit through [our] consciousness'. He claims that, although in some cases these condemnatory 'flashes' may be 'embedded in full-blown emotions', they are typically much more subtle than this in their feeling (Haidt 2012, 45). According to Haidt (2012, 45), moral intuitions typically feel no more emotional than where we can just 'see, instantly' that it would be better to save the lives of five strangers than to save just one.

Here, we might suggest that saving five people rather than one just feels right, in something like the way that 'involuntary' just feels right when we initially look at the letters 'involnutray'. Haidt argues that experiences such as the former are experiences of moral intuition, and he explicitly connects this thesis to the moral sentimentalism of Hume, Hutcheson, and Adam Smith:

> *[M]oral intuition* can be defined as the sudden appearance in consciousness of a moral judgment, including an affective valence (good-bad, like-dislike), without any conscious awareness of having gone through steps of searching, weighing evidence, or inferring a conclusion. Moral intuition is therefore the psychological process that the Scottish philosophers talked about, a process akin to aesthetic judgment: One sees or hears about a social event and one instantly feels approval or disapproval. (Haidt 2001, 818)

Haidt defines (reflective) moral reasoning as follows:

> *[M]oral reasoning* can… be defined as conscious mental activity that consists of transforming given information about people in order to reach a moral judgment. To say that moral reasoning is a conscious process means that the process is intentional, effortful, and controllable and that the reasoner is aware that it is going on. (Haidt 2001, 818)

Haidt's evidence for his moral intuitionism is largely empirical, although he also appears to endorse some of Hume's philosophical arguments (Haidt 2001, 816). Much of the empirical evidence for moral intuitionism came from studies in which Haidt (2001, 817)

presented people with scenarios about actions that were 'offensive yet harmless', such as using a national flag to clean a toilet or masturbating by using a dead chicken. Most respondents immediately judged these actions wrong, but were then unable to explain or to justify their judgements. They often initially claimed that the actions *were* harmful, somehow. When they were convinced that the actions were not harmful, they were often 'morally dumbfounded': they would 'stutter, laugh, and express surprise at their inability to find supporting reasons, yet they would not change their initial judgments of condemnation' (Haidt 2001, 817). Moreover, their 'affective reactions to the stories (statements that it would bother them to witness the action) were better predictors of their moral judgments than were their claims about harmful consequences' (Haidt 2001, 817). What matters here is not the considerations about judgements of harm, but simply that these moral judgements closely correlated with affective reactions. They appear to have all the distinguishing features of intuitive judgements, and none of those of reflective judgements.

In Chapter 8, I will consider the plausibility of moral intuitionism in more depth. In §7.3, I will argue that, granted some independently plausible psychological theories, MFT, along with moral intuitionism, suggests at least one unifying feature of moral judgements. MFT requires that, for any action or motive to be morally salient to us, it must be appropriately related to one (or more) of the relevant psychological foundations. It also suggests that each of us must learn which actions and motives to morally praise or condemn. We should therefore ask whether, for example, we learn to call certain actions 'wrong', or certain motives 'bad', via a particular form of learning. I will argue that this is very plausibly the case, so that paradigm moral judgements comprise a distinctive set of intuitively produced, habitually learned responses to various kinds of socially unacceptable actions and motives. This gives us a unifying feature of moral judgements.

Nothing like this suggestion is considered by Sinnott-Armstrong and Wheatley. While we would need empirical research to demonstrate the truth of this thesis, I will argue that it represents a serious possibility. In Chapter 8, I will argue that moral judgements form a unified psychological kind, and I will develop a more detailed account of the relevant kind of moral learning.

### 7.3. MFT and moral learning

Haidt and Joseph (2008, 367) argue that 'morality might be partially innate, by which we simply mean organized, to some extent, in advance of experience'. Whatever the psychology of our occurrent moral intuitions, our moral development relies not only on our psychological foundations, but also on learning and cultural understanding (Graham et al. 2009, 1030-1031).

These foundational psychological systems 'provide parents and other socializing agents the moral "foundations" to build on as they teach children their local virtues, vices, and moral practices' (Graham et al. 2009, 1030). Haidt and Bjorklund (2008, 205) describe each foundation as 'a "learning module" – a module that generates a multiplicity of specific modules during development within a cultural context (e.g., as a child learns to recognize in an automatic and module-like way specific kinds of unfairness, or of disrespect)'.

MFT does not rest on any complex theory of modularity, such as Fodor's (1983). Instead, 'module' is used fairly loosely, to refer to some bit of 'mental processing' that is 'to some degree module-like', in that its operations are to some extent independent of knowledge 'contained elsewhere in the mind' (Haidt and Bjorklund 2008, 205). For example, 'knowing that two lines are the same length in the Müller-Lyer illusion does not alter the percept that one line is longer' (Haidt and Bjorklund 2008, 205). Merely to the extent that seeing the illusion is psychologically independent of knowing that the lines are of the same length,

Haidt and Bjorklund consider these psychological outputs to be produced by distinct modules.

It is useful to see the kinds of processes that MFT envisions occurring within moral learning modules:

> [I]f there is an innate learning module for fairness, it generates a host of culture-specific unfairness detection modules, such as a "cutting-in-line detector" in cultures where people queue up, but not in cultures where they don't; an "unequal division of food" detector in cultures where children expect to get exactly equal portions as their siblings, but not in cultures where portions are determined by age (Haidt and Joseph 2008, 379).

The claim that each foundation is a learning module may suggest six (or more) distinct ways of learning to moralise. However, given the deliberately loose meaning of 'learning module', it is also compatible with the claim that there is one way of learning to moralise, and six (or more) ways in which an action or character trait may become morally salient, or apt for such learning. This suggests, in turn, something like the following thesis: If, and only if, a person intuitively and negatively responds to a foundational type of action or character trait (where a 'foundational' type is just one that is apt to activate one or more of the relevant psychological systems), *and* that person has learned to treat such actions or motives in some culturally appropriate, negative manner, then the intuition will be, or will cause, a judgement that the action or motive is morally wrong. For example, if a person intuitively and negatively responds to a token case of someone destroying their national flag, then they will judge the action itself as *morally* wrong only if their intuitive response is

connected to their having learned, in some particular way, to treat the destruction of the national flag morally negatively.

If this is correct, then all and only paradigmatic moral judgements are produced in at least broadly the same way: by an intuitive process that occurs in response to an action or character trait that the judger has learned to respond to in some (as yet undefined) way. This is compatible with the claim that different moral intuitions are related to different foundational areas of concern. Presumably we must allow for some variation, possibly including potential exceptions to the paradigmatic norm, given the complexities of human psychology. Nevertheless, this appears to provide a sufficiently unifying feature of moral judgements, as understood by Sinnott-Armstrong and Wheatley.

For this to be plausible, we must allow that each of us can distinguish those of our intuitions that are produced in this way from all other intuitive responses toward actions or motives. There appears to be no phenomenological feature common to all and only moral judgements (Sinnott-Armstrong 2008; Sinnott-Armstrong and Wheatley 2014). However, beyond morality, *many* intuitive reactions occur to us as affective responses of some kind, and may influence our behaviour accordingly, even where the relevant feelings remain unnoticed by conscious reflection (Gendler 2008a; 2008b; Haidt 2001; 2012). We have seen that many intuitively produced judgements are 'micro-valent': they involve motivationally influential but barely noticeable feelings, which cause us to behave in consistent, habitually developed ways towards similar objects (Lebrecht et al. 2012). In such cases, we often recognise that we *are* responding in similar ways towards similar objects, without consciously recognising any phenomenological similarities between these responses. Although there is much that we still do not understand about this, we should allow that moral intuitions may be distinguished from other intuitions even if they are not phenomenologically unified.

Granted this claim, MFT suggests a hitherto unappreciated form of unifying feature for moral judgements: paradigmatic moral judgements share a unifying feature in virtue of the kind of intuitive processes involved in their proximate causes. They are, or they are caused by, all and only those intuitions directed toward objects that we have learned to respond to in some particular way, where we can learn to respond to objects in this way only if they are appropriately related to at least one foundational area of concern. This thesis requires an accompanying theory of the kind of learning involved in moral development.

Of course, the kind of moral learning that I have in mind is a Humean, associative kind. Astute readers will have noticed that micro-valent intuitions bear more than a passing resemblance to Hume's 'calm' passions. MFT is heavily influenced by Hume. As I interpret Hume's metaethics, he can support MFT yet further, via arguments that we judge token actions as morally right or wrong just where we have learned to associate the relevant type of action with causing happiness or unhappiness to those around us. Here, I consider how the essential features of such a theory could cohere with MFT to suggest a plausible account of moral learning.

For Hume, as for MFT, moralising is a fundamentally social kind of behaviour. We learn to approve or disapprove of certain motive kinds by witnessing the effects of such motives, via the actions that they typically produce, on the people who possess them or those around such people. We have seen that Hume's theory of moral judgement is strongly influenced by cases (reminiscent of those used in the empirical studies that led to MFT) in which we judge harmful actions as virtuous, as where we repay a loan to a 'miser, or a seditious bigot' (T 3.2.2.22, SBN 497). Putting aside the many technicalities of Hume's argument, his thesis is that, wherever we have come to associate certain motive types – like the desire to repay loans – with the happiness of those around us, we consistently, habitually approve of any further tokens of these types. Moreover, this consistency in attitude produces

sufficient consistency in behaviour for us to develop and generally adhere to various social systems, such as the generally honest use of money, obedience to governments, and so on.

Haidt (2012, 206) argues that, in the evolutionary development of morality, a crucial role was played by our 'ability to learn and conform to social norms, feel and share group-related emotions, and, ultimately, to create and obey social institutions, including religion'. Apart from Haidt's belief in the moral importance of religion – something that Hume assiduously ignores – the two theorists suggest a very similar understanding of the social pressures that caused, and that may be managed by, the human tendency to moralise.

Here, we may ignore the complex psychological theories underlying Hume's account of moral judgement. Consider merely his claim that we learn to treat certain types of action and character trait as morally wrong just where we associate them with causing unhappiness to those around us. If MFT is correct, then we come to treat certain action and motive types as morally wrong because we have evolved innate but malleable psychological systems that dispose us to intuitively evaluate them in some negative manner. Combining the two theses, we can suggest that typical moral development involves a sensitivity to the attitudes and feelings of people around us, along with a tendency to take as morally salient those cases where foundational types of action and motive are associated with certain positive or negative attitudes or feelings. Further, we can suggest that wherever any kind of foundational action or character trait becomes associated with causing, *inter alia*, anger or distress to those around us, we learn to treat it as morally wrong. This coheres neatly with MFT's view of foundations as 'the psychological systems that give children feelings and intuitions that make local stories, practices, and moral arguments more or less appealing' throughout their moral development (Graham et al., 2009, 1031).

This suggests, for example, that where a person judges a token action wrong just because she thinks it disloyal, she has learned to judge actions of that type as wrong by

associating such actions with some kind of harm or unhappiness, very broadly construed, to those around her. This may be the case even if the action is in not related to her Care/harm foundation in the appropriate way for it to be judged wrong *because* it is harmful.

Now, note Haidt and Bjorklund's (2008, 183) general psychological claim that 'some things are easy to learn (e.g., fearing snakes), while others (fearing flowers or hating fairness) are difficult or impossible'. For example, we find it hard to develop the intuition that we ought to treat everyone equally (Haidt 2012, 182). In contrast, the Loyalty/betrayal foundation readily facilitates the development of intuitions that desecrating a symbol of one's nation is wrong (Haidt 2012, 140). Perhaps only a few experiences of such actions causing anger or distress to members of one's own social group are required. However, common experience of, for example, parts of the UK demonstrates that, where our community remains unperturbed by such actions, we are unlikely to judge them as wrong. It therefore appears to be only where action types are both foundational *and* suitably associated with upsetting people around us that we learn to experience moral intuitions towards them.

The input of one or more of the moral foundations, at some developmental stage, appears to be psychologically necessary to make any action or character trait morally salient to us. However, once any action or character trait *is* made morally salient, we may nevertheless learn to treat it in moral terms in the same way as with every other action or character trait, and via a form of learning that is unique to the moral domain. At least, this suggestion is consistent with MFT.

Consider Haidt's (2012, 3-4) story about someone masturbating with a dead chicken before eating it. Liberals were unique in their responses to this, in that they 'frequently ignored their feelings of disgust and said that an action that bothered them was nonetheless morally permissible' (Haidt 2012, 96). MFT's best explanation for this is presumably that, although many liberals *are* disgusted by the performance of sexual acts with dead chickens,

most typically display very little displeasure toward people who perform disgusting but harmless actions. At least, any displayed displeasure seems to be too weak, or perhaps not censorious enough, to easily produce the kinds of associations that cause us to intuitively judge such actions as morally wrong. Therefore, paradigm wrongness intuitions require associations of some such kind. They presumably just *are* those intuitions that are directed towards some foundational type of action or motive that is associated, by the judger, with causing some relevant kind of displeasure to people around her. Therefore, they all share this unifying feature.

In the next chapter, I will argue that moral judgements do not merely share a unifying feature in this way but that, more strongly, they form a unified psychological kind. For now, I will conclude this chapter by briefly suggesting how even the weaker thesis discussed here may help answer an otherwise puzzling question: Why do people frequently, but not always, mention harm when they are pressed to explain or justify their moral judgements, even where the relevant actions are clearly harmless?

### 7.4. The puzzle of moral dumbfounding

We saw, in §7.1, that De Villiers-Botha (2020), DeScioli et al (2012), Gray et al. (2014), and Schein and Gray (2015; 2018) all argue, in different ways, that people only judge actions wrong if they think that they are harmful. If MFT is correct, as I assume here, then people often judge token actions wrong which they do not think are harmful. Nevertheless, Gray et al. (2014, 1609) present a strong argument against MFT, suggesting that people may intuitively evaluate actions as harmful even where they reflectively believe otherwise. They note that, if Gendler (2008a) is correct, then such clashes of intuition and reflective belief occur reasonably frequently. Moreover, they suggest that this kind of clash explains why people often cite harm as a justification of their wrongness judgements, even where no harm

appears likely. For example, Haidt (2012, 24) notes that, when he asked people to evaluate stories which were 'carefully [written] to remove all conceivable harm to other people', 38% of 1,620 respondents 'claimed that somebody was harmed'.

This objection from Gray et al. certainly suggests that the relevant moral judgements involved some kind of intuitive association between actions being judged as wrong and harm, even where the judger reflectively believes the action to be harmless. Indeed, a further study, developed to test a very similar claim, found that 'participants who condemned [ostensibly victimless] behaviors as wrong perceived a victim 89% of the time' (DeScioli et al. 2012, 145).

To repeat, I do not intend to provide detailed objections to the claim that all wrongness judgements are directed towards token actions that are thought to cause or risk harm to some (real or fabricated) victim. However, we should note the tendency of such theories to rely on unusually broad definitions of 'harm' and 'victim'. For example, Schein and Gray (2018, 47), require a definition broad enough to encompass 'damage to the body, the future self, the soul, or society'. Using a somewhat narrower definition of 'harm', we might, consistently with MFT, explain why people frequently think of harm when they judge clearly harmless actions to be wrong.

This is best illustrated with an example. Even as a self-identifying liberal, I would disapprove of keeping money that was accidentally deposited in my account, even if I knew that nobody could be harmed. If I contemplate doing this, I find that I am disposed to think not only of my own guilt, but also of the disappointment, distress and anger of those people who are important to me. These thoughts are broadly thoughts of harm, although by a narrower definition of 'harm' than Schein and Gray's. Just as Hume argued that we habitually think of harm whenever we contemplate stealing, even where we know that no one would be hurt, I appear to intuitively think of harm to the people that I care about when I

think of performing this action. This suggests only an *indirect* relation between my wrongness judgement and thoughts of harm: one grounded in my moral learning, and in my associations between this action kind and negative reactions from my social circle. However, if people often think of harm when moralising in this way then, when they want to justify their moral judgements and in the absence of any other available reason, it seems quite plausible that many of them would cite harm as a justifying reason for their judgement. This is not, of course, to vindicate their doing so, but merely to suggest a potential explanation for the phenomenon.

In this chapter I have, aside from setting out the details of MFT and moral intuitionism, argued that moral judgements very plausibly share a common, unifying feature, in virtue of the habitual, associative manner in which learn to moralise. I have argued that all and only moral judgements are likely to be produced via a certain kind of intuitive process, given this understanding of moral learning. Moreover, I have argued that, if we learn to judge actions as wrong because we associate them with causing harm or distress to those around us, then this may well explain why people often cite harm when pressed to explain or justify their judgements that token harmless actions are wrong. In Chapter 8, I will suggest a theory by which all and only moral judgements are of a certain psychological kind: moral alief.

# Chapter 8. The Moral Alief Theory

> Utility is only a tendency to a certain end; and were the end totally indifferent to us, we should feel the same indifference towards the means. It is requisite a *sentiment* should here display itself, in order to give a preference to the useful above the pernicious tendencies. This sentiment can be no other than a feeling for the happiness of mankind, and a resentment of their misery; since these are the different ends which virtue and vice have a tendency to promote (M App1.3, SBN 286).

In this chapter, I will combine some core elements of Hume's theory of moral judgements, notably his thesis of Generality (as set out in §3.1.2), with Gendler's account of 'alief', to provide a novel account of the metaphysics of, and causes of, moral judgements. I will argue that every paradigmatic moral judgement is caused by a thought or perception (in the typical, non-Humean sense of the word) of an action or character trait of a kind which the judger generally associates with causing happiness to others (for positive moral judgements) or harm to others (for negative moral judgements). Here, 'happiness' and 'harm' are to be understood very broadly. By this account, moral judgements are, or include, occurrent aliefs. I will leave open the possibility that paradigmatic moral judgements are something *more* than occurrent aliefs: my key claim is that they are at least occurrent aliefs.[1] It is in virtue of these aliefs that

[1] Given the complexity of human psychology, we must allow for at least the possibility that some moral judgements are distinct from aliefs. What follows suggests that, if such judgements occur, they are rare.

we possess 'a feeling for the happiness of mankind, and a resentment of their misery': they play the metaethical role of Hume's moral sentiments.

The resulting theory of moral aliefs builds on Kriegel's (2012) argument that, although some moral judgements are beliefs, others are aliefs. Kriegel (2012, 470) argues that moral beliefs, but not aliefs, feel as though they are about 'objective' matters of fact, whereas moral aliefs, but not beliefs, are directly motivating. I argue, more strongly, that moral judgements are typically moral aliefs. The moral alief theory therefore entails moral intuitionism, as discussed in §7.2. This is because, as Kriegel (2012, 474) argues, Gendler's occurrent aliefs 'play the theoretical role' of the products of our associative, intuitive psychological system(s). In other words, if Gendler's general theory is correct then intuitions just are aliefs.

Given the above, I do not claim that the moral alief theory is entirely original: many of its core theses are already well supported by the work of Kriegel and Haidt, among others. However, I aim to build on their research and arguments, primarily by providing an account of the *content* of moral aliefs, via a consideration of their causes.[2] To do so, I will build on Hume's thesis of Generality.

The resulting moral alief theory states that, for any paradigmatic moral judgement, at least three key psychological states are in play: the *moral judgement* itself – an occurrent alief – which is caused by a *thought or perception* of an action or character trait, where this causal relation should be understood by reference to a background, associative state of the kind which Gendler (2008a, 645) calls a '*dispositional alief*'. For negative moral judgements, relevant dispositional aliefs associate types of action and trait with the harm, unhappiness,

[2] I stress that nothing in this chapter is intended as an account of the semantic content of moral utterances.

distress, pain or displeasure of others (hereafter, simply 'harm'). The cause of any negative moral judgement is thus a thought or perception of an action or trait of a type which the judger *generally* associates with harming others. For positive judgements, relevant dispositional aliefs associate types of action and trait with pleasing or benefitting others. The cause of any positive moral judgement is a thought or perception of an action or trait of a type which the judger *generally* associates with causing beneficial consequences for others.

Gendler understands aliefs as distinct from beliefs. I concur, although what follows is compatible with the claim that alief is a subset of belief. Whichever view (if either) is ultimately accepted, aliefs typically motivate actions, regardless of one's (other) beliefs. If aliefs are distinct from belief, therefore, then my theory is potentially incompatible with the 'Humean' theory of motivation, which asserts, *inter alia*, that motivation necessarily involves relevant beliefs (e.g. Smith, 1994; Sinhabubu 2017).[3] In what follows, where I talk of 'beliefs', I intend to refer to only those beliefs that are not aliefs.

In §8.1, I consider the details of Gendler's account of alief. In §8.2, I argue that, if we understand paradigmatic moral judgements as aliefs, then we can easily explain cases where we disapprove of harmless token actions of generally harmful types. In §8.3, I argue for five further reasons to endorse the resulting theory of moral judgements. §8.4 addresses what is surely its most serious challenge: that of coherently and plausibly explaining the role of reflective thought and reasoning in the formation of many moral judgements.

[3] Kriegel (2012, 480) provides a compelling argument to suggest otherwise, by noting that a moral alief may 'involve a cognitive attitude towards one content and a conative attitude towards another, where the attitudes are modally separable'. This seems compatible with a broad form of Humeanism and with the moral alief theory. However, the moral alief theory can afford to be neutral on the truth of Humeanism, so I leave this topic here.

### 8.1. Aliefs

Consider the motivation behind the following action:

> **Debauchee**: Sally owes money to a friend, who has recently been spending all his spare money on alcohol. She believes that both she and her friend will be better off if she does not return the money, and she would like to keep it. However, she judges that it would be wrong to do so, so she returns the money.

We understand that Sally returns the money because of her moral judgement that it would be wrong not to. Here, I offer answers to two questions about such judgements: What *are* moral judgements? What are their causes?

Hume discusses a situation like Sally's when he asks why we are motivated by moral duty to perform some actions, such as repaying a 'profligate debauchee', which we believe will cause no happiness to anyone (T 3.2.1.13, SBN 482). I will call these kinds of actions – ones we judge as wrong, and which we take to be harmless tokens of typically harmful types – 'harmless wrongs'. In Chapter 3, we saw that Hume sees a parallel between our disapproval of harmless wrongs and those feelings of fear that we may experience when we are high up but believe ourselves to be safe, as when suspended in a 'cage of iron' (T 1.3.13.10, SBN 148).

With this in mind, consider the motivation behind the following action:

> **Iron Cage**: David is suspended from a high tower in an iron cage and is appreciating the view. He believes himself perfectly safe, but he is frightened, so he leaves.

David leaves the cage despite his belief in his personal safety and his desire to remain. He acts on a conflicting desire to leave which he experiences because he judges (at least in a loose sense of the word 'judges') his situation unsafe.

Gendler discusses a similar case, concerning people who feel a fearful reluctance to step onto the Grand Canyon Skywalk – a glass walkway some 4,000 feet above the floor of the Grand Canyon – even though they 'wholeheartedly *believe* that the walkway is completely safe' (Gendler 2008a, 635, Gendler's emphasis).[4] These people, like David, appear to somehow judge their situation unsafe. Gendler argues that we should not understand this last kind of judgement as a belief, and that beliefs play no role in our motivation to flee in such cases.[5] Instead, the role typically ascribed to belief is taken by an alief: 'a mental state with associatively linked content that is representational, affective and behavioral, and that is activated – consciously or nonconsciously – by features of the subject's internal or ambient environment' (Gendler 2008a, 642). David's alief causes him to act *as if* he believes himself unsafe.

David's fearful reaction is an occurrent alief, comprising a mental representation (*R*), an affective response (*A*) and a behavioural response (*B*). Gendler (2008a, 635) understands

[4] Gendler (2008b, 562) also notes Hume's discussion of the iron cage. She follows Loeb (2005) in attributing to Hume the claim that the man in the cage experiences contrary beliefs. In Chapter 3, I argued that we should think of at least one such putative belief as a 'quasi-belief'.

[5] I shall not attempt to explain the distinction between motivation and (mere) causation here, but simply stress that David's alief has desire-like content, and that it is likely to cause reflective desires. This may allow us to say that he has a motivating reason to flee the cage (e.g. Schroeder 2007; Smith 1994). Even if not, David's alief may certainly cause him to act to achieve a goal, so he is surely motivated in some minimal sense (e.g. Alvarez 2010).

an alief like David's to have 'roughly the following content: "[*R*:] Really high up, long long way down. [*A*:] Not a safe place to be! [*B*:] Get off!!"'.

Gendler also describes dispositional aliefs:

> A subject has a *dispositional alief* with representational-affective-behavioral content *R*-*A*-*B* when there is some (potential) internal or external stimulus such that, were she to encounter it, would cause her to occurrently alieve *R*-*A*-*B*. (Gendler 2008a, 645)

Dispositional aliefs are so constituted that, whenever a subject experiences a perception or thought with representational content *R*, the relevant occurrent alief occurs *automatically*: 'without the intervention of conscious thought' (Gendler 2008b, 557).

David's occurrent alief can only be understood by reference to a dispositional alief, which has been shaped by previous experiences over his lifetime, of thoughts or perceptions of people being suspended high up, immediately followed by thoughts or perceptions of them falling to their injury or death.[6] These experiences have produced in David a dispositional alief such that he associates being high up with injury and death. Therefore, wherever David perceives himself as high up, a mental representation of being high up will automatically activate, alongside a fearful affective response and an aversive behavioural response. These are associated such that the former psychologically cannot occur without the latter.

Aliefs include, but are not limited to, (some) paradigmatic emotional reactions, like David's fear reaction. For some aliefs, however, the affective component may be much less

[6] It is possible that, although many dispositional aliefs are produced by experience, some may be hereditary. If it transpires that the fear of heights is hereditary, the example of David's alief may be transposed, *mutatis mutandis*, to any learned fear.

phenomenologically noticeable. For example, Gendler (2011, 44) argues that someone who chooses to invite a person with a stereotypically white name to a job interview, rather than a similarly qualified person with a stereotypically black name, may do so because they alieve that white people are more competent than black people. This racist alief is unlikely to involve any observable, emotion-like feeling. Equally, an alief's behavioural component may sometimes merely ready the aliever to action, without producing noticeable movement, as when a frightened person acts bravely. Gendler intends 'alief' to cover a broad range of automatic judgements, not just obviously emotional responses like David's.

### 8.2. Moral aliefs

For Hume, to judge something morally wrong is to experience a calm sentiment of disapprobation towards it. We saw, in Chapter 3, that he believes that disapprobation is always caused via unreflective, associative, and automatic processes. It is caused by a quasi-believed idea of a character trait, where the trait is considered merely as a token of some general type, and where that type of trait is one that we generally associate with causing harm. As we saw in Chapter 2, Hume at least sometimes calls these kinds of generalised ideas, 'general notions'. We also saw, in §3.3, that disapprobation may cause a calm aversion: a desire to avoid or to demote such traits.

We in the 21st century should, I suggest, similarly understand a paradigmatic moral wrongness judgement as being, or including, a *disapproval alief*. What Hume calls 'disapprobation' is, very roughly, similar to the affective component of an alief. The kind of calm aversion that Hume understand to be caused by disapprobation is, very roughly, similar to the behavioural component of an alief. And, I suggest, a general notion is, very roughly, similar to the representational component of an alief. Indeed, we saw in §3.1.1 that Hume believes that what we would now call 'implicit bias' is due to quasi-believed ideas about the

characteristics of people of certain groups. These ideas are caused by unreflective, associative, and automatic thought processes – general rules – and they may cause us to entertain 'a prejudice' against people, 'in spite of sense and reason' (T 1.3.13.7, SBN 146-7). Without wanting to overstate the similarities between this and Gendler's (2011) theory of implicit bias aliefs, which are certainly founded on very different assumptions about human psychology and reason, there are undoubtedly some interesting parallels between Hume's account of 'general rules' and Gendler's account of the processes that produce occurrent aliefs (T 1.3.13.8, SBN 147).

Of course, many moral judgements involve no obvious feelings. However, we saw in Chapter 7 that Gendler is not alone in arguing that affective responses are frequently involved in motivation, even where they go unnoticed. Hume's concept of calm passions appears at least somewhat vindicated.

If moral judgements are, or include, aliefs, then we must ask what the content of any such alief might be. We have seen that Gendler describes an alief like David's as having '*roughly* the following content: "Really high up, long long way down. Not a safe place to be! Get off!!"' (Gendler 2008a, 635, emphasis added). However, she also claims that any creature that is 'capable of responding differentially to features of its environment that impinge upon its sensory organs has aliefs' (Gendler 2008b, 558). Therefore, she cannot assume that aliefs have linguistic content, for such content is unavailable to many creatures with aliefs. We should, I think, understand her meaning as follows. 'Really high up, long long way down' specifies the content of a non-linguistic representation of being high up. This is accompanied by what Gendler (2008a, 643) understands as 'the experience of some affective or emotional state' (*A*), and 'the readying of some motor routine' (*B*). For David, *A* is a feeling of fear ('Not a safe place to be!') and *B* is a psychological readying to leave ('Get off!!').

Now consider that human dispositional aliefs must be able to represent not just token experiences, such as that of being high above the ground, but also general *types* of experience. For example, David might be told that he will, at some future point, have an experience which will be in some unknown way similar to that of being suspended in a cage, and the thought of this might activate a further, similarly fearful, occurrent alief. Our evolutionary past has, presumably, integrated our propensity to form aliefs with more recently evolved propensities, such as those involving the use of language and conceptual thought. Certainly, conceptualised types of situation or event – such as 'being high up' or even 'perilous situation' – must sometimes be represented within aliefs, if only because we sometimes respond fearfully to thoughts of these, just as we do to experiences like David's.

What, then, would be the content of the *R* component of a disapproval alief? Here we should note that Sally presumably disapproves of keeping her friend's money just because she thinks that doing so would be stealing. The stimulus for Sally's moral judgement is thus the thought of keeping the money owed to her friend, where this action is conceptualised by her as 'stealing'. Following Hume, I suggest that it is only because Sally *generally* associates actions of this conceptual type with causing harm that she disapproves of this token action. Of course, this is not to claim that wherever Sally associates any action with causing harm, she will deem it wrong all things considered. I return to this point in §8.4.

Presumably, from a very young age, Sally – like everyone – has developed and refined a conceptual framework of actions, so that she readily takes a wide range of token actions to be of certain general types. If Sally sees someone bend their knees and propel themselves upwards, she conceptualises the action as 'jumping'. If Sally sees someone take an object that she understands to belong to someone else, she conceptualises the action as 'stealing'.

Also, from a very young age, Sally – like everyone – has witnessed and heard about many token actions that have caused harm to others, and many token actions that have benefitted others. Over time, she has witnessed strong correlations between actions conceptualised into certain categories (like 'kindness') and beneficial consequences. Equally, she has witnessed strong correlations between actions conceptualised into different categories (like 'stealing') and harmful consequences. Like all of us, as part of her moral education, she has been told repeatedly that actions of these types generally have such effects. Many real and fictitious examples have been impressed on her. All this will have caused Sally to associate some action types with the property of benefiting others, and other action types with the property of causing harm to others.

If Hume is broadly correct then, wherever these kinds of associations are sufficiently strong (and, I am assuming, provided that the action types are appropriately related to a foundation, as discussed in Chapter 7), Sally will be disposed to experience approval towards any token action of the first general type, and disapproval towards any token action of the second general type, regardless of her beliefs about their particular consequences. Thus, she disapproves of harmless wrongs.

In many or all cases, therefore, action *types* must be represented within dispositional aliefs, where sufficient tokens of such actions have caused harm, so that any future action which is taken to be of the same type will also activate disapproval. The action that Sally contemplates in Debauchee is not thought by her to cause harm, and it does not obviously resemble any action which is considered harmful, except insofar as it is conceptualised as an act of stealing. It is disapproved of *because* it comes under the concept 'stealing'. Therefore, we should take it that a representation of the action type 'stealing' is included within Sally's dispositional disapproval alief.

This putative dispositional alief is such that the thought of stealing automatically produces an occurrent alief with content *R*-*A*-*B*: roughly, 'Stealing. Guilt! Avoid!!'. We can generalise from this example to the claim that every judgement that one's own action is or would be wrong is, or includes, an alief with similar content, where *R* represents the action type in question. If one judges that someone else's act of stealing (for example) is wrong, we can take it that the content of the alief would be roughly, 'Stealing. Blame! Punish!!'

One reason to think that moral judgements are aliefs is that this thesis explains why we disapprove of harmless wrongs. Aliefs are produced by automatic processes, which perform a different psychological role from that of (paradigm) beliefs. The habitually acquired, automatic nature of an alief means that it cannot be easily resisted or altered according to the available evidence on any token occasion. To return to our parallel with a fear alief, wherever one has come to associate heights with danger, one cannot easily avoid feeling fear when high up. In contrast, Gendler (2008b, 565-566) argues that beliefs *aim* to 'track truth', by being 'evidence-sensitive': 'The man suspended in the cage *believes* that he is safe because if he were to gain evidence to the contrary, his attitude would change accordingly'. However, we know that he cannot, and so will not, quickly change his alief in response to evidence. Gendler thus understands beliefs to be evidence-sensitive in a way that aliefs are not. A belief that one is in danger will typically change to (or be replaced by) a belief that one is safe, even if one still appears to be suspended high up, if one has strong evidence, aside from this appearance, that one is safe. In contrast, a fear alief like David's will not easily alter so long as one appears to be suspended high up, no matter what additional evidence one has of one's safety.

Nevertheless, many aliefs are to some degree evidence-sensitive, even if not in the same way as beliefs. David's dispositional alief is sensitive to the evidence that being at a great height generally causes harm. Implicit racist biases are sensitive to racist messages in

society, which falsely indicate that black people generally have negative properties (Gendler 2011). Aliefs can *only* be generally sensitive, because of the automatic and psychologically necessary relation between perceiving relevant stimuli and experiencing the relevant thoughts and attitudes. The *R* components of dispositional aliefs represent object types where these are associated with certain types of event, and this is sufficient to activate an occurrent alief in response to *any* token of the relevant object type, whether the associated event is believed to be imminent or not.

I therefore understand Gendler to mean by the 'evidence-sensitivity' of a judgement that the judgement is sensitive to the evidence regarding token situations, as with David's belief that the token experience of being high up is safe, because he is in the cage. This belief is sensitive to evidence about the strength of the cage and so on, whereas his alief is not. Sally's disapproval alief is insensitive to the evidence that, in the circumstances of Debauchee, an act of stealing would benefit all concerned. However, it is sensitive to the evidence that, generally, stealing causes harm.

Because everyone has different experiences, we are liable to differ to some extent in the types of actions which we associate with harming or benefiting others. Some of this diversity of moral judgement is explored in detail by MFT, as we saw in Chapter 7.

This diversity of judgement is consistent with the moral alief theory. For example, in conservative cultures any subversion of authority is likely to be deemed morally wrong, whereas in more liberal cultures such actions appear morally neutral, perhaps even laudable (Haidt 2012, 142-144). Moral concerns about subverting authority are made salient by the Authority/subversion foundation. In conservative cultures, actions that challenge authority clearly cause more unhappiness than they would in more liberal cultures, where they are more readily accepted. Children who grow up in conservative cultures are therefore much more likely to associate such actions with harm (in our broad sense of the term) than their

liberal counterparts would. Presumably, the representations within a person's moral aliefs may, in many cases, be caused by the *moral* attitudes of parents and others. All that is required for someone to judge an action wrong is for them to possess a disapproval alief, such that they associate the type of action in question with causing harm, where 'harm' includes unhappiness and distress. If one's parents strongly disapprove of an action type, then their disapproval may be sufficient to *cause* that association, particularly if they are visibly upset or offended by any reference to such actions.

Indeed, someone brought up in an ascetic household may frequently witness other people's disapproving or angry attitudes towards generally pleasurable action types, and so associate such actions more with these attitudes – and with the unhappy consequences of these attitudes – than with their directly pleasing effects. In many such cases, it might be that the original non-moral reasons for these action types being the cause of distress are forgotten by all concerned, although the moral distress remains.

Nevertheless, there is clearly much consistency across moral judgements, especially across wrongness judgements. Most people, whatever their cultural background, are distressed when they see other people being physically or psychologically harmed. MFT is consistent with the claim that most moral judgements in most cultures are directed towards actions that are seen to be physically or psychologically harmful in some way (Haidt et al., 2015). The moral alief theory adds to this that all paradigmatic moral judgements are at least indirectly related to harm, broadly construed: they are directed towards actions of types that tend to produce unhappiness or distress within the judger's cultural context. This is not to say, of course, that this relation would be cited by the judger in any justification of her judgement.

Given this complexity and more, all I can provide here is the beginnings of the moral alief theory. Nevertheless, one reason to endorse it is that it explains our tendency to disapprove of harmless wrongs. In §8.3, I argue for five further reasons to endorse the theory.

## 8.3. Five arguments for the moral alief theory

### *8.3.1 Motivation*

The moral alief theory readily explains the frequent connection between moral judgements and (defeasible) motivations to act in ways deemed appropriate in light of these judgements. This connection is apparent even to those who deny motivational internalism: the view that moral judgements are necessarily motivating, at least to some degree (e.g. Smith 1994). Whatever the truth of this thesis, moral judgements are typically closely related to motivation, and this must be explained. Kriegel (2012, 478) argues that all aliefs are 'inherently' motivating, although Gendler (2008a, 644) allows that some atypical aliefs may not be motivating. If moral judgements are aliefs, therefore, then they will be at least typically motivating, and perhaps necessarily so.

### *8.3.2 Intuition*

The process by which an occurrent alief unfolds is, we have seen, an intuitive one: it is automatic, associative, and unavailable to consciousness. We have also seen that Haidt (2001; 2012) has argued that many moral judgements are produced intuitively, partly by showing that people frequently form such judgements rapidly, and then struggle to justify them.

Consider too Hume's claim that 'vice and virtue are not matters of fact, whose existence we can infer by reason' (T 3.1.1.26, SBN 468). He argues that, if we reflect on why we form any one moral judgement, then we will be unable to explain any reasoning that led to the judgement, because we do not use reason to infer that the action is wrong. His

argument suggests that, if people are asked why they take actions of the type called 'murder' to be wrong, then they will generally be unable to provide any satisfactory reasoning for this, so that the judgement that murder is wrong cannot be a reasoned belief. This is presumably to be contrasted with cases where people make – and *can* satisfactorily argue for – similarly widespread non-moral judgements, such as the judgement that most birds can fly. Hume would certainly allow that *this* is a reasoned belief, because we can clearly identify and report on the ideas and reasoning that we employ when coming to this judgement.

Of course, people may give reasons why they take actions to be wrong, but such reasoning generally rests on further moral claims, as when someone argues that eating meat is wrong because it causes unnecessary harm, and that causing unnecessary harm is wrong. In such cases, a full explanation of why one judges eating meat to be wrong would include an explanation of why one judges causing unnecessary harm to be wrong. Hume's argument suggests that people generally cannot explain these kinds of foundational moral judgements. This seems plausible: ask a range of people if it is wrong to harm or kill others for personal gain and most will assent, but ask them to explain why it is wrong and most will struggle to do so in any detail. While some philosophers and psychologists will give (frequently contradictory) detailed reasons, most laypeople will be unable to do this. They may refer to relevant facts, like the painful effects of such actions, but most will be unable to provide any supporting premises to argue from such non-moral facts to their conclusion. Press them and they will likely be dumbfounded: they will simply assert that some types of things are wrong, while realising that they cannot give reasons for this.

If moral judgements are typically formed via reflective processes, then we should expect to be able to explain the reasoning behind these judgements in some detail, as we can for reflective beliefs like 'most birds can fly'. However, if they are produced automatically, without involving conscious attention, then the relevant processes will be unknown to us, and

we should not expect to be able to provide detailed explanations of our reasons for forming them.

Haidt's research demonstrates that many moral judgements are intuitive. Hume's argument suggests that even our most universally agreed and foundational moral judgements are intuitive. Therefore, the most parsimonious theory of moral judgements will explain them as all being formed, in large part, by intuitive processes of some relevant kind. I will deal with a pressing objection to this conclusion – resulting from the prevalence of reflective reasoning in moral discussion – in §8.4. For now, note that the moral alief theory meets this requirement.

### *8.3.3 The Puzzle of Moral Dumbfounding, revisited*

In Iron Cage, we assume, David will not simply feel fear and run. He will try to make sense of his fearful response via reflective consideration. Although his fear alief activates just where he perceives himself to be high up, and it involves no representation of falling, he will surely entertain thoughts of falling to his death. Reflecting on these thoughts will help him explain his behaviour to himself.

This kind of interplay between alief and other forms of thinking seems typical: for humans, where any alief represents object type O because O is associated with event type E, then any occurrent alief activated by a token of O will typically cause a thought of E. While a dog suspended in a cage may, perhaps, simply fear being high up, because this is all that the relevant alief represents, David's is a fear of falling as well as of heights, because thoughts of falling swiftly follow his fear.

Something similar is suggested by Gendler's (2011) argument that implicit racist biases are aliefs. Gendler never details what she takes the R-A-B elements of racist aliefs to be, but she does allow that they cause thoughts of – to use her example – black people as

lazy, poor, criminal and so on (Gendler, 2011, 43). Indeed, without these thoughts of complex properties, such aliefs would be merely affectively negative and aversive towards black people. If we are to explain why implicitly racist people frequently judge black people as criminal, for example, we must allow that implicit bias aliefs consistently produce, or otherwise accompany, consciously available and relevant thoughts. Such thoughts may be relatively simple, intuitively-produced ones – perhaps of a single word like 'criminal' – but without them these aliefs could not cause the kinds of racist effects which we know implicit biases cause. To give an example, a racist alief may have roughly the following content: 'White person. Unwelcoming! Avoid!!', and it may engender a thought of the white person as hostile.

Therefore, if the moral alief theory is correct, then we should predict that people will typically think of broadly harmful consequences (E) when they judge an action (O) wrong, even where they believe that the action is harmless. Haidt's (2001; 2012) research, which we considered in Chapter 7, suggests that this is the case. I have already argued that something like the moral alief theory can explain what I called the 'puzzle of moral dumbfounding': our propensity to cite harm when we judge actions wrong, even where there is clearly no risk of harm. I now argue that the moral alief theory *predicts* something like this phenomenon.

Recall that, in Haidt's studies, people were asked to morally evaluate actions, each of a type that generally would cause harm or distress, but each of which is described in the scenario as causing no harm or distress. For example, Haidt (2012, 38) tells a complex tale of incest in which all risk of harm is removed: nobody other than the siblings know about it, they are secure from psychological harm, pregnancy is avoided and so on. Approximately 80% of people judge this action to be wrong, even where they acknowledge that no harm is involved.

We saw that, when people were asked to justify their disapproval of the actions in these scenarios, 38% of respondents explicitly claimed that they were harmful (Haidt 2012, 24). This is a surprisingly high percentage of people who claim that harm is present in a short scenario precisely designed to exclude it. As Gray et al. (2014, 1609) argue, the fact so many people cited harm suggests that they intuitively associate harm with the wrongness of the scenario. Moreover, it seems likely that more than 38% of respondents entertained *some* thought of harm, given that many who did so would surely not mention it as a reason, because they would realise that it had been specifically excluded.

Haidt's research therefore indicates that, even where people do not believe that an action which they deem wrong will cause harm, many of them are disposed to think of it as being harmful, and to relate this thought to their reason for judging it wrong. This relation between judging something harmless as wrong and thinking of it as harmful requires an explanation. In Chapter 7, I argued that any broadly Humean, associationist theory of moral learning could provide such an explanation. Here, I claim that, given that aliefs typically operate such that any occurrent alief activated by a token of object type O will typically cause a thought of event type E, we should *predict* that judgements of moral wrongness will typically corelate with thoughts of harm, broadly construed. It is therefore unsurprising if at least some people cite harm in their explanations or justifications of such judgements, even if their judgements are directed towards clearly harmless token actions.

### *8.3.4. The unity of wrong actions*

The moral alief theory is well placed to explain and, to some extent, vindicate a common intuition: that all and only wrong actions share some feature or property in common, aside from their being the objects of our disapproval.

It is certainly tempting to think that *something* unites those objects which we call 'wrong', beyond their all being the objects of our disapproval.[7] As Jackson et al. (2000, 94) argue, it is a 'platitude' that non-moral similarities and differences in acts are 'relevant to moral similarities and differences in acts'.[8] We act as if we can demonstrate the wrongness of actions by referring to relevant non-moral features.

This behaviour may seem perplexing, however, when we consider the research underpinning MFT, and the many kinds of actions which may be judged wrong. These actions seem to form, as Prinz (2007, 48) puts it, a 'hodgepodge' of different kinds, seemingly united only by our disapproval of them. This apparent disunity is partly why so-called 'Cornell realists' argue that 'wrongness' must refer to something like a 'homeostatic cluster' of properties, rather than a single property (Boyd 1988, 196). Indeed, McDowell (1981) claims that no one non-moral property *could* be shared by all and only wrong actions because, if it were, then someone could potentially identify this property and thereby grasp the extension of 'wrong': all without sharing or even understanding the moral concern. McDowell is surely not alone in finding this prospect highly implausible.

Nevertheless, I think that Jackson et al. (2000, 87) are right to argue that, because we use the predicate 'wrong' to mark a distinction between wrong actions and other actions, *something* must unify wrong actions by distinguishing them from others.

Consider McDowell's worry first. As Roberts (2011) argues, we may simply find the prospect of identifying a unifying, non-moral property of wrong actions implausible because the relevant property is no part of the meaning of 'wrong'. Therefore, no reductive analysis of

[7] I assume that this claim, and relevant arguments, can be applied, *mutatis mutandis*, to other moral terms.

[8] More strongly, Blackburn (1984, 187) argues that it would be 'absurd' to deny that two or more 'naturally identical states of affairs compel the same moral description'.

the term is possible, so that it is highly unlikely that anyone could successfully grasp the extension of 'wrong' without moralising. Perhaps the property exists, but research is required to discover it. If so, and given our assumption that we seek a naturalistic theory of wrongness, then any plausible and comprehensive metaethical theory must include an account of a unifying, natural property of wrong actions, where this property is no part of the meaning of 'wrong'. Note that this requirement avoids potential problems from Moorean open question arguments (Moore 1903).

It is unlikely that any non-disjunctive, non-moral, natural property is *intrinsic* to all and only wrong actions, or we would presumably have observed this property by now. It is, however, entirely possible that all and only wrong actions share a non-disjunctive, non-moral *relational* property, so that certain actions – each with quite different intrinsic properties – affect us in some way such that we disapprove of all and only these actions, and thus call them 'wrong'. Note that this is purely a claim about the psychological causes of moral judgements. As such, it is potentially compatible with some forms of moral realism, like Cornell realism, but also with other, very different, metaethical views like subjectivism and expressivism.

Of course, the relevant relational property cannot simply be that of causing disapproval, because we are seeking this property to explain *why* only some intrinsic properties of actions make us disapprove of these actions. Further, this property must be something of which we are largely unaware via reflection or introspection, otherwise we could easily categorise wrong actions in non-moral, relational terms (as if, to give an implausible example, we called all and only actions which sadden us 'wrong'). This consistent lack of reflective awareness is independently plausible, given that moral judgements appear to be intuitively produced: it is very likely that the relevant relational property is unavailable to conscious attention. We form wrongness judgements without

realising precisely why, and we therefore struggle to identify the unifying property of wrong actions. However, if we are even dimly aware that something unites those actions that we call 'wrong', aside from their being the objects of our disapproval, then this could certainly explain why we expect all and only wrong actions to share one or more non-moral properties.

Therefore, we should expect that, for any moral judger S, all those actions that S sincerely judges by negative (or positive) moral judgements are so judged because they have a non-disjunctive, non-moral, naturalistic, and relational property in common, which causes S to disapprove (or approve).

This expectation is met by the moral alief theory. The relational property of any action which results in it causing disapproval is that of being intuitively associated by the judger with causing harm to others. All and only such actions are judged wrong, although this is clearly not what we mean by calling an action wrong. Moreover, this is consistent with the claim that we judge different actions as wrong for different reasons; indeed, we may do so for fundamentally different *kinds* of reasons, as MFT suggests. Nevertheless, as we saw in §8.3.3, we appear to be influenced in our conscious thinking by our associations between all those actions that we call 'wrong' and harm. This suggests that we *do* appear to be at least dimly aware that something unites those actions that we call 'wrong', aside from their being the objects of our disapproval.

*8.3.5 The fundamental importance of harm to wrongness*

Here, I present a brief argument that is closely related to the one in §8.3.3, that the moral alief theory predicts something very like the phenomenon that I called the 'puzzle of moral dumbfounding'. I want to build on my claim, in §8.3.4, that we frequently assume that something unites all wrong actions beyond our disapproval of them. The moral alief theory is ideally placed to reconcile two highly plausible but jointly perplexing claims: that people

often judge harmless actions to be wrong, and that many of us assume that all wrong actions are somehow related to harm. Even those who argue against this assumption note the temptation to make it (e.g. Prinz 2007, 48; Haidt 2012, 4). Indeed, if we did not assume a relation between wrong actions and harm then we would feel no need to explain our disapproval of harmless wrongs.

The moral alief theory states that we all associate all wrong actions with harm, so that we are disposed to think of them as harmful. I suggest that this may plausibly allow us to explain why many people assume that only harmful actions are wrong, despite the obvious existence of harmless wrongs. Where we are disposed to think of each token of some type of action as harmful, as we do with each token morally wrong action, it is surely unsurprising if many of us assume that the type itself is closely related to harm.

### 8.4. Moral reflection, reasoning, and belief

I have argued that all paradigm moral judgements are occurrent aliefs. Yet occurrent aliefs unfold purely via unconscious, intuitive processes, whereas many moral judgements are formed only after much reflective reasoning and debate. According to social intuitionism, most moral reasoning is post hoc, and primarily undertaken with the aim of persuading oneself and others that one's moral judgement is correct (Haidt & Bjorklund, 2008, 189-190). Yet even Haidt (2001, 819) suggests that *some* moral judgements may be reflectively produced. Kennett and Fine (2009) argue that relatively few moral judgements are intuitive, because there is evidence that people often override their moral intuitions by reflective reasoning, and so form judgements which oppose these intuitions. How can we reconcile the fact that moral judgements frequently occur after significant moral reasoning and deliberation with the claim that these judgements are aliefs, given that paradigm aliefs occur without any reflective thought?

Kriegel avoids this problem by arguing that many moral judgements are beliefs. He cites as evidence the phenomenology of moral judgement, which, he claims, 'often involves a feeling as of homing in on an objective matter of fact' (Kriegel 2012, 470). In fact, recent psychological evidence suggests that most people sometimes talk as if this were so, but at other times they clearly allow that moral judgements are subjective responses (Goodwin and Darley 2008; 2012; Pölzler and Wright 2019). This is consistent with Kriegel's view, for it may be that people take pure moral beliefs to be about objective matters of fact and pure moral aliefs to be subjective responses.

However, I argued in §8.3.2 that the most parsimonious theory of moral judgements would explain them as all being formed, in large part, by intuitive processes. Hume and Haidt have each, I suggested, provided good reasons to think this is the case. Moreover, if Gendler (2008b, 554) is correct, then aliefs are extremely prevalent in our thinking of all kinds, even if we only notice them where they conflict with our reflectively produced beliefs. I therefore suggest that all typical moral judgements are aliefs, although we may often form complexes of reflective, non-moral belief *and* moral alief.

Here, I can only provide the basis of an account of how such complexes might be formed, but I believe this demonstrates that a plausible account of moral reasoning *can* be reconciled with the moral alief theory. A clue lies in the complex roles which moral aliefs play in our lives, compared to most aliefs. Loosely put, the function of a fear alief, for example, is to provide a quick and powerfully motivating aversion to anything which generally causes harm. Moreover, this is how we typically understand the role of such fearful reactions. Therefore, where one feels fear but simultaneously believes that one is safe, one's fearful reaction is seen as mistaken. In Iron Cage, David understands his reaction as a fear of falling and, as such, he takes it to be misrepresentative.

Moral aliefs are atypical because they do not simply enable us to manage individual situations, but to manage the interpersonal coordination of actions over time. It is sometimes noted, as by expressivists like Stevenson (1963) and Blackburn (1998), that a crucial advantage of moralising is that it allows for a coordination of behaviour between people with different interests. Moral discussion allows us to agree – to some extent – on which actions are to be encouraged, permitted or discouraged. I suggest that this is achieved via the automatic activation of moral aliefs whenever tokens of relevant action types occur. By forcing us to be consistently influenced by the *general* effects of actions like stealing, moral aliefs cause us to generally converge in our disapproval of such actions, even in many cases where the effects do not otherwise displease us. In this way, these aliefs strongly assist the coordination of our actions. If each of us reacts with at least some disapproval to the idea of a token act of stealing, whatever we each take to be its likely effects, then this will ensure that we are all inclined, to some degree, to abstain from and demote it. Moral aliefs therefore help us to agree on what is to be done.

Here, my argument partly follows that which I take Hume to be employing in his discussion of a common point of view, as discussed in Chapter 6. At least within one culture, moral aliefs – somewhat like Hume's moral sentiments – are to some extent *uniform*, in that they typically respond in similar ways towards all tokens of any one type of action, regardless of how the particular token actions otherwise affect us (for example, stealing almost always causes disapproval). I do not think they are anything like as uniform as Hume assumes, because of the clear extent of moral diversity even within any one culture or society. Nevertheless, there is much moral agreement, and, to this extent, there are many action types towards which many of us will experience similar moral aliefs.

Moral aliefs therefore allow us to converge on our evaluations of actions in a way that we could not easily do otherwise. Consider an act of shoplifting, for example. Perhaps those

who know the shoplifter and the difficulties she faces will judge that, as the supermarket owner will hardly suffer, this is an understandable and forgivable act. The supermarket head of security, however, faced with many cases of shoplifting and aware of the cost of these, is likely to judge that each theft is part of a greater harm, and must be punished. In many such cases, it would be difficult to agree on the degree to which an action should be punished, and similar ones discouraged, if each of us judges only from what Hume calls 'our particular and momentary situation' (T 3.3.1.23, SBN 587).

Hume argues from this, as we have seen, that the coordination of our actions requires 'some other standard of merit and demerit, which may not admit of so great variation' (T 3.3.1.18, SBN 583). This seems correct; we require a means of judging actions which generally causes us to agree, and which generally supports us in negotiating our various, and often conflicting, desires and interests within a complex society. This cannot be easily achieved by considering only our non-moral preferences on particular occasions. However, if we all broadly agree, at least within the confines of any one culture or society, on which *types* of actions are to be permitted, discouraged or promoted, then this helps us to agree about the best ways to treat token actions in particular situations. Focusing on our moral aliefs helps us to agree in this way. We have seen that, once an action is conceptualised as a token of a relevant type, like 'stealing', then a moral alief automatically follows. If this occurs for all of us in at least roughly the same way, then we will at least have the basis of a means to agree about how best to respond to any token act of stealing.

Of course, we may have very different moral priorities, by this or any plausible theory of moral judgement. One person may feel greater disapproval towards stealing than she does towards social injustice, whereas another person may feel the opposite. Here, the former person would be likely to treat an impoverished shoplifter more harshly than the latter person would want to allow. Nevertheless, both people may still be in broad disagreement that,

insofar as the action is one of stealing, it is wrong to some extent, even though it is also an *injustice* – and so also wrong – that anyone in a wealthy society should be required to endure poverty. They have something of a shared evaluative framework within which their debate may proceed.

A further complication is due to the fact that it is not always a straightforward matter to decide how best to conceptualise any token action. This is particularly the case since we may conceptualise an action without applying a clear name or label to it. For example, an act may surely be conceptualised as one that is 'unlikely to cause harm' or 'understandable in the circumstances', just as much as it may be conceptualised as 'stealing'. Some people may experience disapproval aliefs towards actions conceptualised as 'harmful', 'subversive', 'disgusting' or even 'other than that action that is likely to cause the greatest amount of happiness in the circumstances'. These are all plausible conceptual categories for actions to fall under, and, for all I know, we may also mentally categorise actions into kinds that we have no terms for. Each type of conceptual category may produce different occurrent moral aliefs, to different degrees of strength.

Moreover, there seems no reason why one person could not possess all the above kinds of intuitive association. I will argue that the interplay between these as we consider any one action can explain much of our moral indecision, reasoning and argument. As Haidt and Bjorklund (2008, 191) note, we do not typically engage in moral reasoning by calmly asserting reasons, but instead aim to 'trigger the right intuitions in others', by making emotive claims. We use reflective reasoning to try and cause intuitive judgements in others which, we hope, will make them see things our way.

The primary type of argument involved in moral reasoning is, if I am right, a form of what Stevenson (1963) calls 'persuasion', rather than empirical reasoning or logical analysis. These latter types of reasoning may often be involved in morally relevant decision-making,

of course, as where we want to decide on which of our available actions causes the least harmful outcome. However, moral disagreement *as such* typically involves a focus on persuading others, primarily by the use of affective language, to agree with us in their moral judgements. We may attempt to persuade our opponents to apply a 'laudatory title' to an action of which we approve, or a disparaging title where we disapprove, so that we may influence their affective response to it (Stevenson 1963, 44). A friend of Sally's may say, 'it's a kindness to keep the money', or 'it would be cruel to let him buy more alcohol'. In this way, he tries to make her question the type of action under discussion, so that she responds to the token action differently. If he can persuade her to reconceptualise the action from a type with a negative moral association to a different type with a positive moral association, then her disapproval of the action is highly likely to wane.

Consider how moral reasoning very often proceeds. If we want to make someone approve of an action, we may describe it in moral or non-moral terms, but we almost always use terms associated with benefitting others. Similarly, we typically use (moral or non-moral) terms associated with harm to persuade people to disapprove of an action, even if our reasons for our own moral judgement do not rest on the claim that it is harmful. The anti-abortionist is *pro-life*: preserving life is generally very beneficial. He may call his opponent *uncaring*, content to *kill* and *murder*. His opponent is *pro-choice*, and she stresses the harm caused to women by denying their freedom to choose. Abortion is a *right*, but also a *medical procedure*: something with strongly positive associations in a world where medicine helps many. Each disputant tries to make the other (or, perhaps more realistically, any undecided listeners) conceptualise abortion in terms which activate the relevant moral aliefs for their view.

Of course, any action may be conceptualised into many distinct categories simultaneously, but in moral disagreement we aim to make some categories seem more

salient than others. If Sally is persuaded that, despite being in one sense an act of kindness to keep her friend's money, when all is said and done it would still be *stealing*, then this may cause sufficient intuitive disapproval for her to reflectively endorse it, so that her overall judgement is that the action is wrong. Her disapproval may still be countered by some degree of approval, given that the act of stealing is, in this case, a *kind* and *helpful* one. Nevertheless, stealing is very strongly associated with harm, so that persuading someone that an act is best considered as an instance of stealing is highly likely to lead to an overall judgement that it is wrong.

Sally's reflectively endorsed judgement may be a very complex one, involving, among other things, a belief that the action is (best conceptualised as) an act of stealing.[9] This allows us to address one worry about the moral alief theory: that moral judgement seems to us more like belief than like a paradigm alief such as a fear alief. We often experience moral judgement as, to use Horgan and Timmons's (2006, 263) phrase, 'a matter of psychologically "coming down" on whatever issue is under consideration'. This is much the same way that we experience the process of coming to form a belief after deliberation. I think we can explain this by reference to the fact that a moral judgement will often be formed only after much discussion and reflective reasoning, as beliefs about matters of fact often are. Moreover, as with many kinds of alief, a moral alief typically involves no obvious feeling. It is, therefore, unsurprising if its phenomenology is similar to that of a belief about a matter of

[9] Perhaps Sally's moral judgement sits within a complex that comprises further beliefs too. For example, if she is both a moral realist and a rule utilitarian, she might believe that the action under consideration instantiates a property of wrongness, and that it is prohibited by a rule which, when uniformly adhered to, causes the greatest possible happiness. It is an interesting question how these beliefs might relate to her overall judgement that the action is wrong, but I do not pursue this question here.

fact.[10] Nevertheless, according to the moral alief theory, Sally's reflectively endorsed *moral* judgement is a disapproval alief: it is only in virtue of this that she judges that the act of stealing is morally wrong.

In this way, moral intuitionism is consistent with the claim that moral reasoning may be effective in changing our minds, and for good, justifiable reasons. I agree with Horgan and Timmons (2007, 282) view, that moral 'reason-giving' can, on Haidt's view, appear to be nothing other than 'confabulation', and that we should resist this notion.[11] Although we should allow, as Haidt does, that moral reasoning always follows moral intuition, we should also see this as a frequently iterative process. We often consider reasons for viewing token actions and characters in different ways, and we are likely to come to different moral judgements about them, depending on how we end up understanding them. Haidt's 'theoretical stance' may entail, as Horgan and Timmons (2007, 291) suggest, that 'the phenomenology of reason-based appropriateness [of some moral judgements] is a systematic illusion', but we need not alter it much to remove this worry. There are more or less appropriate ways of categorising actions and characters, and moral reasoning involves reasoning about these.

According to Haidt (2001, 819), almost all genuine moral reasoning is performed in inter-personal discussions and debates. It should be clear that I agree that inter-personal

[10] This is, of course, very similar to Hume's argument that, because the moral sentiments are 'calm', they 'are very readily taken for the determinations of reason' (T 2.3.3.8, SBN 417).

[11] Horgan and Timmons (2007, 286-287) argue for what they call 'morphological rationalism' which is, very roughly, the view that our moral principles 'guide' the formation of our moral intuitions, so that our reasons for these principles indirectly inform our moral intuitions. This allows that any post hoc reasons are likely to be good ones, but it does not allow, as my view does, for genuine reasoning to occur during moral deliberation about particular cases.

reasoning is of particular importance to us (and I will say more about this in Chapters 9 and 10). However, I stress that this type of reasoning may also occur alone, where we deliberate about the best way to think of an action or character. I am not convinced that we only rarely engage in moral reasoning to question our own moral attitudes, as Haidt believes.

The moral alief theory therefore allows, as any plausible theory of moral judgements must, that our moral practices involve much reflective reasoning and debate, of which much but not all concern considerations of harm. It also allows that moral reflection and argument often produce beliefs about the nature and effects of actions, and about how they should therefore be conceptualised. Further, these beliefs may cause us to overturn our moral judgements, as Kennett and Fine stress. Yet, however long and reflective the chain of reasoning is which leads to any moral judgement, we have seen very good reasons to think that the judgement itself will be, insofar as it is a *moral* judgement, an intuitive one. Once Sally has finally determined that the action is best conceptualised as stealing, then she disapproves, no matter what evidence she has about the actual effects of the action on people's happiness. Her moral judgement is thus insensitive to evidence. To this extent, it is an alief.

I cannot, of course, hope to develop a fully satisfactory and detailed account of moral reasoning in one chapter. However, I believe that I have suggested the basis of one plausible account of moral reasoning that is compatible with the moral alief theory. Moreover, I hope to have shown that the moral alief theory is a viable one, although I have only been able to provide the basis of what must surely be a very complex theory here. I strongly suspect that further empirical work will be required to determine the plausibility or otherwise of this theory, but for now I present it as a viable option, deserving of further consideration.

In Chapter 9, I will consider how this theory of moral judgements might best cohere with a semantic account of our verbalised moral evaluations.

# 9. Emotive subjectivism

> [W]hen you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it… And this discovery in morals… is to be regarded as a considerable advancement of the speculative sciences; tho'… it has little or no influence on practice. Nothing can be more real, or concern us more, than our own sentiments of pleasure and uneasiness; and if these be favourable to virtue, and unfavourable to vice, no more can be requisite to the regulation of our conduct and behaviour (T 3.1.1.26, SBN 469).

The moral alief theory may tell us something about the nature of typical moral judgements or moral reasoning, but it says little about moral language. If it is true, then what do we mean when we call an action, motive or character 'wrong', for example? In this chapter and the next, I present one potential answer to this question. I argue, along with others like Barker (2000), Copp (2001; 2009), Finlay (2014), and Strandberg (2012), that what we imply when we use moral language is as important to us as are the meanings of our moral utterances. I suggest a subjectivist theory of the meanings of moral terms, alongside a more complex account of the pragmatics of moral language. As Hume stresses in the passage above, although the meanings of moral terms are simple, we use moral language to 'regulat[e]… our conduct and behaviour', which is undoubtedly a complex practice. In honour of its emotivist forebears, Hume and Stevenson, I call my theory 'emotive subjectivism'.

In Chapter 5, I argued that Hume's is an emotivist theory. His arguments, particularly those related to what I called his 'Vivacity Thesis', are too intricately connected to his theory

of ideas to persuade us now. Nevertheless, I will argue for a theory of moral language based, in large part, on the emotivism of Stevenson (1963), along with the 'quasi-realist' expressivism of Blackburn (1993a; 1998).[1]

Here is an outline of key claims from Chapters 7 and 8. Paradigmatic moral judgements are aliefs: intuitive, learned responses towards actions (and character traits) of types that we associate with causing happiness or harm, broadly construed. Moral aliefs play an important role in helping us coordinate our actions and attitudes within a society. When we engage in moral reasoning or disagreement, we typically try to determine the best way to categorise any one action or trait, so that we will experience what seems to us an appropriate moral alief. Our reflectively reasoned and endorsed moral judgements are, at least as a rough approximation, those moral aliefs that we experience when we consider an action or motive as falling under what seems to be the most appropriate morally salient category, among the potential categories under consideration.

If something like this theoretical view is accurate, then moral judgements are presumably important to us primarily *because* they help us to decide, and to agree on, what to value and how to act. Should we consider a token act as 'stealing' or as 'kindness'? If we are persuaded to adopt the former categorisation, then we are at least typically *ipso facto* persuaded to disvalue the action and to refrain from or discourage it. If the latter, then we are at least typically *ipso facto* persuaded to value it and to perform or encourage it. We can say something very similar about *types* of action: if we are persuaded that abortion is best thought of as a *medical procedure that empowers women* rather than as a form of *murder*, then we are

[1] Whereas emotivists argue that moral utterances express emotions, perhaps broadly construed, expressivists argue that they express motivating attitudes of some kind(s), which might be very different from emotions.

likely to experience approval towards abortion, and so to be at least somewhat moved to value it and to perform or encourage it.

The most detailed and sophisticated extant theories of moral language that focus on its practical nature are expressivist theories, like Stevenson's and Blackburn's, along with others such as Gibbard's (2003) and Horgan and Timmons's (2006).[2] These endorse the 'Humean ontology' of morality: an anti-realist ontology such that morality is to be explained purely by reference to human attitudes and to the ordinary, naturalistic features of the world towards which such attitudes may be directed. Other theories that endorse this ontology include Dreier's (1990; 2009) speaker relativism, and dispositional theories, such as Egan's (2012), Lewis's (1989), and Prinz's (2007). I will have more to say about these in Chapter 10, but for now I will assume that the account of the practical nature of moral language embedded within expressivism (although perhaps not unique to it) is broadly correct. I take it that some such account provides our current best way to understand the practical nature of moral language, within the framework of the Humean ontology. Indeed, I do not deny that an expressivist theory might ultimately be preferable to the theory that I develop here. However, I have at least two reasons for pursuing emotive subjectivism over an expressivist theory.

The first of these is that the most detailed and persuasive forms of expressivism, including Blackburn's and Gibbard's (2003, 148), rely on a deflationary account of truth that we may not want to accept. It is, I think, worth our while to consider how we might draw on, in my case, Blackburn's careful treatment of moral practice without having to rely on deflationism about truth. One of my aims will be to vindicate ordinary moral talk and behaviour, as Blackburn aims to, but without requiring any one theory of truth. This may

[2] Horgan and Timmons argue for a form of 'cognitive expressivism', by which beliefs may be either descriptive or normative in nature, and by which moral utterances express normative beliefs.

have a further benefit. It is often claimed that expressivism has a problem with understanding how simple moral sentences, which have no descriptive meanings, can be meaningfully embedded in more complex sentences, where they appear to require descriptive meanings. The resulting 'Frege-Geach' problem is seen by many as insurmountable (e.g. Schroeder 2008). My own view is that Blackburn's (1998) theory has the resources to overcome it. However, we can note that moral terms are straightforwardly descriptive, according to emotive subjectivism, so that the problem does not arise for this theory.

My second reason is that the moral alief theory suggests a different picture of moral thinking than those presented by current expressivist theories. Indeed, the picture of moral thinking that I reject is implicitly accepted within most metaethical theories, including speaker relativism and dispositional theories. The moral alief theory – and, we will see, several recent psychological studies, aside from Haidt's – suggest that moral thought and language is significantly less reflective, and appears significantly less consistent, than most metaethicists generally assume.

In asking what moral utterances mean, metaethicists have typically assumed that, as Gill (2009a, 216) said just over ten years ago, 'ordinary [moral] discourse is uniform and determinate enough to vindicate one side or the other' of various metaethical disputes. Gill (2009a, 232) doubted the veracity of this assumption, and recommended that 'descriptive meta-ethics [should] involve much more empirical investigation than it typically did in the 20th century', to discover how people in fact use moral terms. As we will see in Chapter 10, such studies are now being undertaken, and they generally suggest that the above assumption about ordinary moral discourse is mistaken, as Gill suspected.

Moreover, although it is generally acknowledged that much moral knowledge is tacit, I suspect that metaethicists often overestimate the extent to which people can interpret their own moral judgements. More precisely, given the intuitive nature of moral judgements, I

strongly doubt that introspection on our own judgements of moral wrongness can help much in the process of coming to understand the meaning of moral utterances. Indeed, my theory will rely on the thesis that most of what we think when we judge something wrong is unavailable to conscious attention, as I have argued over the last two chapters. Call this the 'opacity thesis'.

I hope to show that we have good reasons for moralising as we do. The opacity thesis allows for what I will call a 'strongly indexical' account of moral utterances, such that they often refer to the moral aliefs experienced at the time of utterance, regardless of the tense of the moral sentence being uttered. I will argue that this allows for a method of vindicating moral language that is closely related to, and draws heavily on, Blackburn's quasi-realism. For simplicity, I will focus mainly on one type of moral utterance: that of the form '*x* is wrong', where *x* is a token action.[3] I will call this 'MU' (for 'moral utterance').

In §9.1, I provide the essential elements of emotive subjectivism. In §9.2, I argue for a subjectivist theory of the meaning of moral terms. In §9.3, I argue for a pragmatic account of moral language. I will further motivate and defend emotive subjectivism in Chapter 10.

### 9.1. Situating emotive subjectivism

Emotive subjectivism sits within the wider theoretical framework of so-called 'hybrid theories'. These typically aim to combine the most plausible elements of *cognitivism*, by which moral utterances express moral beliefs, and *expressivism*, by which moral utterances express desires or other motivating attitudes. As Schroeder (2009, 258) notes, it appears plausible to many that 'cognitivists avoid expressivists' problems with logic and inference

[3] Following Grice (1968, 226), I employ an 'artificially wide sense' of 'utterance', to cover cases of, for example, writing that some action is wrong.

because they associate moral sentences with ordinary descriptive contents, and… expressivists can offer elegant explanations of the motivating power of moral judgments and the pull of the Open Question argument'. A theory which allows that moral utterances express both beliefs *and* desires may therefore seem attractive.

Fletcher (2014) distinguishes two claims that a hybrid theorist might make:

(i) moral thought: Moral *judgments* have belief and desire-like aspects or elements.

(ii) moral language: Moral *utterances* both ascribe properties and express desire-like attitudes (Fletcher 2014, 173).

In this chapter, I will defend a hybrid theory of moral language, by which moral utterances ascribe properties and *implicate* desire-like attitudes. In §9.3, we will see that this (probably) sits within the class that Fletcher (2014, 173) calls '*implicaturist hybrid* views of moral language'. Not all hybrid views entail anything about implicature; examples of non-implicature views include Boisvert (2008) and Ridge (2014).

Most typical hybrid theories, whether of the implicature hybrid kind or otherwise, claim that any description that occurs when we call something 'wrong' is given by the belief-like element of our moral judgement, and that we simultaneously (semantically or pragmatically) express the desire-like element of that judgement. So, for example, Barker (2000) argues that, when we say, 'X is good', we mean that X has some natural property F, and we imply that we are committed to approving of F things. Here, the moral *judgement* that X is good includes a belief that X has some natural property F, and the moral *utterance* 'X is good' expresses that belief, and so ascribes property F to X.

I do not argue for such close connections between the (somewhat belief-like) representative component of a disapproval alief and the descriptive element of MU. According to emotive subjectivism, we do not express the descriptive component of a moral judgement when we sincerely utter MU, but instead refer *to* our occurrent disapproval alief. Such aliefs, along with occurrent approval aliefs, are what I call 'moral judgements', and so I understand subjectivism as the theory that utterances like MU express judgements about our moral judgements. This is merely a terminological stipulation, of course: I could stipulate that moral judgements just are those mental states, whatever they may be, that we express via utterances like MU. However, I will argue, moral aliefs are of central importance to our moral concerns, whereas our judgements about these aliefs are not.

Roughly following Stevenson (1963), I suggest that an utterance of MU means that the speaker disapproves of *x*, and that the utterance has the expressive function of pressing others to come to evaluative agreement. Here is my preferred formulation of emotive subjectivism:

> A sentence of type MU means that the speaker is experiencing, or would experience, negative (micro)valence towards *x*, of the kind produced by an occurrent disapproval alief. When uttered, this sentence has the pragmatic function of:
>
> (i) demonstrating one's current commitment to disapproval of *x*, and;
>
> (ii) demanding of anyone who does not disapprove of *x either* that they come to disapprove *or* that they provide the speaker with sufficient reason to stop disapproving.

I will generally summarise the literal meaning as 'I disapprove of *x*'. I will sometimes describe the experience of disapproval as a 'feeling', but I do not suggest anything more than micro-valence by this. Unlike Stevenson (1963, 18-19), mine is not an emotivist theory, because I do not claim that any part of the *meaning* of a moral term or utterance is given by what he calls its 'dynamic' use: its being such as to 'give vent to our feelings' or to 'incite people to actions or attitudes'. My theory entails a relatively simple subjectivist theory of the meanings of moral terms, of a kind which Stevenson strongly repudiated in his later work.

Some key points: the features of *x* that cause disapproval may or may not be consciously available to the utterer. As Haidt shows, in many cases, we *feel* that something is wrong, without being able to say why. Very typically, the feeling is accompanied by at least some aversion towards *x*, but I remain neutral as to whether this is any part of the meaning of MU. It may be that some atypical aliefs do not motivate in this way, as discussed in chapter 8.[4] Finally, the meaning of moral sentences can be either *strongly* or *weakly* indexical: if *strongly*, then we mean that we disapprove of *x*, here and now; if *weakly*, then the meaning of the sentence is more closely related to the tense of the sentence. A strongly indexical utterance of '*x* could be wrong, but I doubt it', means that I could be, but doubt that I am, experiencing disapproval towards *x*. A weakly indexical utterance of '*x* could be wrong, but I doubt it', means that I could come to disapprove of *x* in other circumstances, as where I learn more about *x*, but I doubt it.

Admittedly, the suggested meaning of the strongly indexical utterance of '*x* could be wrong, but I doubt it' appears highly counterintuitive. It certainly seems that we can doubt

[4] So-called *amoralists* – people who sincerely express moral sentences like MU without experiencing any motivation to act in accordance with them – are theoretical possibilities by this account (e.g. Brink 1986; Smith 1994, 60-76). If they exist, however, they are rare: aliefs are *very* commonly motivating.

whether an action is *wrong*, but it does not seem to most of us that we can doubt whether we are experiencing disapproval or not. Generally, facts about our own experiences are much more readily available to us than are facts about objects or properties in the world. Indeed, moral facts appear more similar to facts about objects or properties in the world, at least in this respect, than they do to facts about our own experiences. This is one reason why many people find metaethical theories like realism and expressivism more plausible than subjectivism.

However, the opacity thesis is of crucial importance here. Because of their intuitive nature, it is not always clear to us when we are experiencing moral aliefs. There is very strong evidence, some of which has been discussed in Chapters 7 and 8, to show that we are not always aware that we are, or how we are, influenced by aliefs, generally (Gendler 2008a; 2008b). It is, therefore, highly possible on my account that we are sometimes unsure about whether we morally disapprove of an action or not. Perhaps, in such cases, we experience other, somewhat similar, intuitive responses, or we are aware that we disapprove, or not, of other somewhat similar actions. According to emotive subjectivism, at least some cases of moral doubt are best explained as cases of doubt about one's own moral responses.

I will argue that the ambiguity between the strongly and weakly indexical meanings of MU plays an important role in moral language. This is not an ambiguity of which we are always aware, although some psychological studies, to be discussed in Chapter 10, appear to suggest that we are at least sometimes aware of it. Despite this complication, however, mine is a relatively simple subjectivist theory, of the kind that Zangwill (1990, 587) summarily dismisses as '[b]ad old Naive Subjectivism'. Very few philosophers since Hobbes have been tempted by anything similar. In Chapter 10, we will see that there are good reasons for this. Nevertheless, I will argue that, given both the opacity thesis and a suitable account of the pragmatic nature of moral language, subjectivism can and should be defended.

I have only provided a brief outline of emotive subjectivism so far. I will now argue for the subjectivist theory of the meaning of moral judgements, before providing the basis for my pragmatic theory in §9.3.

**9.2. An argument for a subjective account of the meanings of moral terms**

The moral alief theory entails that any paradigmatic wrongness judgement is intuitively produced. There are several aspects of such judgements that we cannot readily come to understand from even careful introspection. For example, we are not consciously aware of the process by which they are produced, and they are learned in a habitual and associative manner that is similarly unavailable to conscious attention. Even where an occurrent alief is, unlike the process which caused it, consciously available, we cannot assume that it is *fully* available, or that we can easily or clearly distinguish it from other, related thoughts and feelings by introspection. Moreover, we typically think of harm when we judge some token action wrong, regardless of our beliefs about it.

This may seem a pessimistic view, given that it not only entails the opacity thesis, but also suggests that much of our moral thinking might be misleading to us. We might wonder whether, whenever we call things 'wrong', we are simply led by quick, intuitive responses to think of them as harmful, even where we reflectively believe that they are harmless.

Consider Sinnott-Armstrong et al.'s (2010) argument, concerning the limits of intuitive calculating power and the implications of this for moral intuitions. In brief, they argue that there is significant evidence to show that intuitive processes cannot perform complex, accurate calculations. These processes can, however, perform fast calculations based on large amounts of unconsciously stored experience, and they often use such calculations to act as *heuristics*. Therefore, Sinnott-Armstrong et al. conclude, if moral questions are complex ones, then moral intuitions are very likely to be heuristics.

To give a bit more detail: when we respond quickly to questions, we often unconsciously substitute a heuristic answer for a carefully reasoned one. For example, if we are asked to quickly estimate the numbers of seven-letter words in a chapter which have 'n' as their sixth letter, and then to quickly estimate the numbers of seven-letter words in the same chapter which end in 'ing', we are very likely to offer a higher number in response to the second question than to the first. However, on reflection we can see that this is impossible. The standard explanation for this type of response is that we rely on an 'availability heuristic': we intuitively substitute the 'target attribute' – the numbers of words of a certain type within a chapter – for the more accessible attribute of the numbers of certain word types of which we can quickly think of examples (Sinnott-Armstrong et al. 2010, 248-250).

Many intuitive processes are heuristic in something like this way. Consider the example of a fear alief, from Chapter 8. A fear alief appears to (roughly) substitute the target attribute of being *dangerous* with the more accessible attribute of being *of a kind that we generally associate with danger*. When we have to make quick decisions about our safety, such responses are often very valuable. However, as in Iron Cage, they can also be misleading.

Sinnott-Armstrong et al. (2010, 268) argue that, according to almost all theories of moral wrongness, moral intuitions can *only* be heuristics, so that they are unlikely to be consistently reliable or to appropriately respond to the target attributes of our moral judgements. For example, act utilitarians believe that an action is wrong only in virtue of its failure to maximise happiness. If they are correct, then the target attribute of a wrongness judgement is that of being *not such as to maximise happiness*. However, moral intuitions cannot then be relied on to accurately respond to wrongness, for this attribute is too complex for intuitive processes to accurately respond to.

The only theories that Sinnott-Armstrong et al. (2010, 257) allow to be exceptions to this are 'those that identify moral wrongness with the judger's own emotional reactions or preferences'. Any other candidate attribute for wrongness, at least given the assumption of naturalism, would be too complex for non-heuristic intuitions to plausibly respond to. Note that this is not a claim about the complexity of potential properties of wrongness, but about the complexity of the thought processes required to correctly identify wrongness, according to most theories of moral wrongness. It applies as much to expressivist theories – at least, to those that entail that we should carefully reflect to appropriately identify wrong actions – as it does to most realist theories.

In §8.3, I argued that whenever we judge something morally right or wrong, we do so because we experience an intuitive judgement – an alief – towards it. Moral aliefs just are the 'emotional reactions or preferences' that cause us, somehow, to utter sentences like MU. Therefore, if we accept the foregoing, we have only three options to explain what we mean when we utter MU: we can deny that there is any attribute in virtue of which anything is morally wrong; we can allow that there is such a target attribute, but accept that moral judgements are about some heuristic attribute; or we can understand MU to mean something like, 'I disapprove of *x*'.

We cannot dismiss the first option. We might be consistently misled into implicitly thinking that *x* has some mysterious but essentially harmful property whenever we utter MU, even where the evidence and our reflective judgements demonstrate that *x* is harmless. This suggestion is reminiscent of Mackie's (1977, 35) contention that the 'basic, conventional, meanings of moral terms' involve a 'claim to objective, intrinsic, prescriptivity'. Like Mackie, I reject any such notion. Therefore, if this suggestion is correct, then I would have to agree with Mackie's (1977, 48) 'error theory', which entails that MU is in all cases false, or at least not true. Perhaps our best option in this case would be to accept a 'hermeneutic

fictionalist' view like Kalderon's (2005), by which we only ever *pretend* that moral properties are ever instantiated.

Fortunately, I do not think we need to accept either an error theory or fictionalism, for at least two reasons. The first of these comes from the opacity thesis itself. We cannot just assume that the semantic content of MU is directly related to the content of a moral alief, because we have reasons to think that there are very complex connections between, on the one hand, the objects of our intuitive judgements and, on the other hand, our conscious thoughts and speech acts. We do not have direct conscious access to the psychological relations between wrongness judgements and harm, and so we should not simply expect moral utterances to be *about* harm. Put simply, MU need not be about what a wrongness judgement is about. Perhaps it is, but we would need to argue for this claim and, I suspect, develop empirical studies to support it. We should not, therefore, assume that MU expresses a moral judgement.

My second reason to reject error theory is more important. As Blackburn (1993a, 149-152) argues, moral anti-realism seems to do little to motivate us to give up moral talk, and any plausible anti-realist theory must explain why this is. Even if we were convinced that all our moral utterances were false, we would surely not simply ignore our moral judgements. We would still be pleased by kindness and appalled by cruelty. We would still experience disapproval towards many harmless wrongs. And, Blackburn (1993a, 150) argues, we would therefore need to develop some language to discuss and to co-ordinate these judgements and attitudes, albeit a language 'purged' of the 'metaphysical mistakes' besetting moral language. Yet there is no reason to think that this new language would be substantively different from our moral language as it currently stands, and so no reason to think that our current language involves the kinds of metaphysical commitments that Mackie suggests it does. Therefore, we

should reject the first option: that there is no attribute in virtue of which anything is morally wrong.

Now consider the second option: that there is a target attribute for our reflective judgements of wrongness, but that moral judgements – moral aliefs – are about some heuristic attribute. If this is right, then we should ask what target attribute the heuristic attribute is replacing. The only plausible candidate is that of causing harm (broadly construed), so that, roughly speaking, wrongness judgements stand to beliefs about harm as fear responses stand to beliefs about danger. Moral aliefs certainly appear to play the psychological role of heuristic guides to questions of well-being, broadly considered.

This appears to lead us directly back to an error theory, such that we mistakenly judge harmless wrongs to be harmful in *some* sense, even as we reflectively believe that they are harmless. However, I acknowledged in §8.4 that moral judgements, broadly speaking, often involve complexes of moral alief and non-moral belief. We should therefore ask whether MU might get its meaning from all or some of the components of such complexes. If so, then we may, perhaps, avoid such an error theory even if we allow that moral aliefs are heuristics.

It is very difficult to know how plausible this might be, given the many possible beliefs that may be relevant as well as the moral alief. Nevertheless, I think we should reject this option. For one thing, there does not appear to be any difference in the *meaning* of MU when it is uttered in immediate response to some action from when it is uttered after careful reflection. But in the first kind of case we presumably have little to go on apart from our intuitive responses, and primarily our moral aliefs, for we will typically not have had time to form reflective beliefs. Moreover, this approach seems to lead to Open Question worries: if moral judgements are at least primarily aliefs, then we can disagree, or be unsure, about any of our reflective beliefs relevant to some action without thereby disagreeing, or being unsure, about whether it is *wrong*. Therefore, MU appears to get its meaning primarily from a moral

alief. If we are to avoid an error theory, we must allow that it does this without being an expression of a heuristic judgement that something is such as to cause harm.

This leaves just one answer: whenever we utter MU, we refer to our own disapproval alief towards *x*. Of course, we do not typically think we mean this by MU. Here, the opacity thesis is crucial to understanding our moral practices (as well as, I trust, keeping us at bay from the dangers of Moorean Open Question arguments). Given the prevalence of aliefs in human psychology, and the intuitive and often micro-valent nature of these, we can very plausibly hold the thesis that we will typically experience a moral alief *whenever* we consider any morally salient object, even where the relevant affective and behavioural effects are not immediately obvious to us.[5] I suggest that this thesis allows for a form of ambiguity about the meaning of moral terms that, I will argue, we employ constructively. Moral terms can often be, but need not be, *strongly* indexical: they can meaningfully refer to one's occurrent moral aliefs, regardless of the tense of the moral sentence. To say that *x* is, would be, was, could be, or will be wrong is, in cases of strong indexicality, to say that *x* would be, was, could be, or will be such that one is, here and now, experiencing negative (micro)valence, of the kind produced by an occurrent disapproval alief, at the thought of *x*.

For example, if I sincerely say, 'even if I stopped disapproving of torture, then torture would still be wrong', then I mean roughly that, even if I were to stop disapproving of torture, then torture would still be such that, here and now, I disapprove of it (I will discuss his further in §10.3). However, this meaning is not obvious to me, because it is far from obvious to me what I am referring to when I call torture 'wrong'.

This semantic interpretation is strongly influenced by expressivist arguments (e.g. Blackburn 1998, 314; Sinclair 2008, 265). I will explain this mainly in chapter 10, but for

[5] In Chapters 3 and 5, I argued that Hume holds this thesis.

now I will simply state my intention to endorse the same general approach to vindicating moral language that expressivists take: to stress that, while it is true that the moral utterances that we make are dependent on our moral judgements, and that our language would change if our judgements were to change, it does not follow that there is anything defective about our current judgements or use of moral language. In particular, understanding this dependence need not mean that we come to focus more on our judgements than on the objects of these judgements, or that we should do so. We moralise because we care, here and now, that (for example) people are not tortured, rather than because we care about how we feel when they are tortured. Moral language refers to some of the feelings by which we care, but, more importantly, it also allows us to express our concern for the victims of torture, among many other things, and to negotiate any relevant practical questions with others.

According to the opacity thesis, our moral experiences, including introspective ones, provide no obvious indications as to whether our moral judgements are caused by our awareness of mind-independent properties (torture is objectively like *that*, and hence appears so to me) or by our tendencies to respond towards certain non-moral properties (torture is subjectively like *that* to me). Whereas Mackie (1977, 35) held that 'ordinary' moral statements 'include a claim to objectivity', I will argue in Chapter 10 that they are typically neutral on such matters: most people employ the predicate 'wrong' without any underlying metaethical commitment. What led Mackie to his error, I suspect, is that we frequently use language that *seems* to entail moral objectivity, or moral realism, and for good reasons. I will have more to say about this in Chapter 10.

Even allowing for my thesis of stronger and weaker indexicality, my theory of the meaning of moral judgements is much simpler than most subjectivist theories. These tend to be dispositional theories, by which MU means that I (some of us, all of us) am (are) disposed to disapprove of *x* under certain conditions (e.g. Egan 2012; Lewis 1989; Prinz 2007). In

Chapter 10, I will argue that, while these can offer much in the way of vindicating moral language, a simple subjectivist account that is coupled with a suitable account of pragmatics can offer more. Before arguing for this, however, I must say something about the pragmatic theory that accompanies my subjectivism.

### 9.3. The pragmatics of moral language

Perhaps the earliest detailed account of the practical purpose of moralising comes from Stevenson (1944; 1963). He saw moral language as fundamentally related to those areas of life where we ask what to value or how to act. As we saw in Chapter 8, he focused mainly on interpersonal disagreement, but he also considered cases of solitary indecision, as where we mull things over or ask ourselves whether something is really wrong. Stevenson's general view of moral language is at least roughly shared by later expressivists, including Blackburn (1993a; 1998) and Gibbard (2003). For example, Sinclair (2016, 2837) notes Stevenson's influence when he delineates moral statements by reference to their societal function, such that moralising is 'a distinctive linguistically infused mutual co-ordination device through which competing parties can negotiate towards (and thence maintain) mutually beneficial and stable patterns of attitudes and actions'. Despite being no expressivist, Haidt (2012, 189-220) suggests that morality evolved so as to fulfil something very like this role, because more morally cohesive societies were better able to survive than less cohesive ones.

As I read him, Stevenson argues for a view in his early work, albeit one which he would later come to reject, by which moral utterances both express *and* describe our moral judgements. For example, Stevenson (1963, 23) claims that calling something 'good' describes one's approval in a way that has the 'dynamic function of giving direct expression' to that approval. Here, to express approval is to utter a sentence that means that one approves, but in such a way that one is thereby seeking to resolve moral disagreement with others, by

trying to influence them to approve, while also showing a willingness to be persuaded to alter one's own attitudes, given sufficient reason.

Stevenson explains his early emotivism by reference to his own, controversial, theory of meaning, which need not concern us here. Shorn of any such entailments, I believe that it provides our best theory of both the meaning and pragmatics of moral language. We therefore need to consider how we might understand pragmatics more generally. I think that the best starting point, at least, is to adopt Grice's (1989) theory of 'implicature', and to claim that MU carries a *generalised conversational implicature*.

### *9.3.1. Gricean implicature*

Grice (1989) argues, very plausibly, that in many cases a complete account of the meaning of the sentences used in any conversational exchange cannot completely account for all that is communicated by that exchange. We very often imply things that are not entailed by the literal meanings of the sentences that we use. He offers this example:

> Suppose that A and B are talking about a mutual friend, C, who is now working in a bank. A asks B how C is getting on in his job, and B replies, *Oh quite well, I think; he likes his colleagues, and he hasn't been to prison yet* (Grice 1989, 24).

As Grice observes, A may well ask what B is implying by saying that C has not yet been to prison, although she may not need to, perhaps if she knows that C readily succumbs to temptation. *Something* has been implied by B's phrase, and we cannot look purely to the meaning of the relevant terms to understand what this is. This gives us an example of what Grice (1989, 26) calls 'conversational implicature'.

There are two kinds of conversational implicature, according to Grice; the general and the particular.[6] The example just given is a form of particular conversational implicature, because it is only due to the particular circumstances in which the phrase 'he hasn't been to prison yet' occurs that the phrase implies anything. *General* conversational implicature occurs where an utterance of some phrase or sentence generally carries an implicature of some kind, unless the implicature is cancelled. Grice's (1989, 37) example is the utterance of 'a sentence of the form *X is meeting a woman this evening*' which generally implicates – at least, if X is a heterosexual man, as Grice seems to be assuming – that X is not meeting his wife or a family member. We can cancel this implication, however, simply by saying that the relevant woman *is* his wife or family member.

Grice (1989, 26) describes our conversations and 'talk exchanges' as 'characteristically, to some degree at least, cooperative efforts [in which] each participant recognizes… to some extent, a common purpose or set of purposes, or at least a mutually accepted direction'. He argues that we all, *ceteris paribus*, expect one another to observe a principle which he calls the 'Cooperative Principle': 'Make your conversational contribution such as required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged' (Grice 1989, 26). This forms the background to our conversations, as we continually assess each other's reasons for saying what we say accordingly.

Grice then sets out four 'conversational maxims', which he claims we generally presume one another to adhere to. Davis (1998, 11-12) summarises these as follows (as cited in Fletcher 2014, 176):

[6] Grice (1989, 25) also discusses a distinct class of *conventional* implicatures, which need not concern us here.

*Quality*: Make your contribution true; so do not convey what you believe to be false or unjustified.

*Quantity*: Be as informative as required.

*Relevance*: Be relevant.

*Manner*: Be perspicacious; so avoid obscurity and ambiguity, and strive for brevity and order.

Conversational implicature typically occurs by deliberately flouting one or other of these maxims. So, to continue with the previous example, by saying only that X met a woman, we are failing to be specific where it would be expected that we *would* be specific if we meant that X had met his wife or family member (Grice 1989, 38). For any successful conversational implicature, we should be able to 'calculate' what is being implicated, by considering what is uttered in light of the four maxims (Grice 1989, 39).

However, we (or, at least, the neurotypical among us) do not usually need to perform any reflective calculation, as we will quickly recognise what is being implicated. General conversational implicature appears to be a prevalent feature of our use of language. Common examples include implying that not all of the essays have been marked by saying that 'some of the essays have been marked' and implying that you cannot have both a drink and a snack by saying 'you can have a drink or a snack'.

A useful, albeit rough, way to think of general conversational implicature is to consider it as what is implicated when we say something that, had we not intended to implicate something, would not have been well said in the way that it was said. To use a different example, of a kind which Grice (1989, 37-38) also discusses, why say 'I was in *a*

filthy house last night' unless I meant to imply that it was not *my* house? If it *were* my house, then I should be inexplicably violating *Quantity* by not saying 'my' house.

*9.3.2. Implicature and moral utterances*

Barker (2000), Copp (2001; 2009), Finlay (2014), and Strandberg (2012) all develop implicature hybrid theories, by which moral utterances have descriptive meanings and imply desire-like attitudes. None of these theorists endorse or consider anything like the moral alief theory, of course, and none are attracted to any form of subjectivism about meaning.

Barker (2000) and Copp (2001; 2009) argue that uttered moral assertions carry what Grice (1989, 25) calls 'conventional', rather than conversational, implicatures. Finlay (2014) suggests something very close to an account by which moral utterances carry particular conversational implicature, although he does not use this language. He develops his own account of the pragmatics of language, albeit one that is influenced by Grice's theory.

To illustrate just how different implicature hybrid theories may be, consider that Strandberg's (2012) is the most similar to mine, insofar as it is the only other theory (to my knowledge) to claim that assertoric moral utterances typically carry a generalised conversational implicature. However, Strandberg's main aim is very different to mine: to reconcile cognitivism with the intuition behind motivational internalism. This is the resulting theory:

> *The Dual Aspect Account (DAA)*: A person S's utterance of a sentence of a type according to which φing has a certain moral characteristic, such as "φing is wrong," conveys two things: **(i)** The sentence expresses, in virtue of its conventional meaning, the belief that φing has a moral property. **(ii)** An utterance of this type of sentence carries a generalized

> conversational implicature, *GCI*, to the effect that S has a certain action-guiding attitude in relation to φing. (Strandberg 2012, 101)

Clearly, Strandberg is unwilling and unable to vindicate moral language while adhering to the Humean ontology, as I hope to do, because he allows for moral properties, as I do not. Given the many differences between mine and other implicature hybrid theories, I do not intend to argue against other such theories directly. Moreover, I do not insist that mine *is* an implicature hybrid theory, as set out here. First, there have been several criticisms of Grice, and we may want to endorse some modified understanding of implicature, such as Davis's (2003). More fundamentally, Fletcher (2014, 195) argues that, while 'GCI' views such as Strandberg's fare better than other implicature hybrid theories, views which endorse a 'Simple Pragmatic Story' may fare even better. This story is as follows:

> (1) People commonly have desire-like attitudes in accordance with their moral judgments.
> (2) People's moral utterances voice their moral judgments accurately, at least for the most part.
> (3) (1) and (2) are common knowledge. (Fletcher 2014, 195)

I see no reason why the essential features of my account should not conform to something like this story as well as to Grice's or Davis's. If so, we would have our relatively simple subjectivist account of the meaning of moral terms, along with a simple pragmatic account by which:

(1) People are commonly committed to adhere to any moral judgement to which they refer, and they typically want others to either come to share that judgement or to convince them to change their own judgement.

(2) People's moral utterances refer to their moral judgments accurately, at least for the most part.

(3) (1) and (2) are common knowledge.

For this reason, although I am happy to consider mine as an implicature hybrid theory, sitting within a Gricean framework, I rely only on the formulation in §9.1, which is neutral regarding theories of implication. What matters more than the theory of pragmatics that we might want to employ is our understanding of what we are doing when we moralise, such that moral utterances have the implications that I take them to have. I address this in Chapter 10, where I will argue that we moralise in something like the way that Blackburn and other expressivists have previously suggested.

# 10. What We Do When We Moralise

> Those who have denied the reality of moral distinctions, may be ranked among the disingenuous disputants; nor is it conceivable, that any human creature could ever seriously believe, that all characters and actions were alike entitled to the affection and regard of every one… Let a man's insensibility be ever so great, he must often be touched with the images of RIGHT and WRONG; and let his prejudices be ever so obstinate, he must observe, that others are susceptible of like impressions (M 1.2, SBN 169-70).

In this final chapter, I consider some of the ways in which ordinary moral language proceeds, and I try to demonstrate that many of the quasi-realist strategies for vindicating it may be applied within the emotive subjectivist framework. Like expressivists, I understand moral language as providing a pragmatic solution to problems of social coordination. As expressivists do, I take it that our use of moral language typically reveals only our first-order moral judgements, rather than any metaethical commitments. Unlike any expressivist, however, I will argue that we have learned to moralise by referring to our moral judgements via language that often, but not always, seems to suggest that we hold some kind of realist view.

Realists are cognitivists who believe that some of our moral beliefs are true. Such beliefs are often taken to be about what Shafer-Landau (2003, 15) calls 'stance-independent' moral facts: facts that obtain, independently of any given actual or hypothetical human perspective, such that some things just are morally good or bad, right or wrong. Some realists, like Brink (1989), offer what Sinclair (2012) calls the 'presumptive argument', which

states that we should presume realism unless proven otherwise, because ordinary people appear to be implicit realists. In contrast, Sinclair, Blackburn and other expressivists aim to vindicate this language, but without requiring that we endorse anything other than the Humean ontology.

My view is strongly influenced by expressivists like Blackburn, Sinclair and, as we have seen, Stevenson. I will argue that some of Blackburn's (1993; 1998) 'quasi-realist' arguments can be employed, largely unchanged, within the framework of emotive subjectivism. In this way, I hope to vindicate ordinary moral practice and language in a very similar way that Blackburn seeks to do. However, we will see some good reasons to think that ordinary moral practice and language is not so realist-seeming as is often thought.

In §10.1, I consider the quasi-realist approach to explaining and vindicating ordinary moral language. In §10.2, I respond to a pressing objection to my theory, which I call the 'incompatibility objection'. In §.10.3, I ask what those features of this language are that require vindication, and I argue that ordinary moral language does not typically imply any metaethical theses. We need only explain why people sometimes *appear* to believe in stance-independent moral facts, nothing more. Finally, in §10.4, I begin the work of vindicating moral language, so understood, by means of emotive subjectivism. I will conclude without having achieved a full vindicatory account, which I believe will take significant work to achieve, but with sufficient reason to be optimistic that such an account is obtainable.

### 10.1. Quasi-realist expressivism

Expressivism, in Blackburn's words, 'says that we *voice* our states of mind, but denies that we thereby describe them' when we moralise (Blackburn 1998, 50, Blackburn's emphasis). The negative thesis here is, of course, a denial that expressivism is any form of subjectivism. Most expressivists, and their emotivist precursors, have stressed this denial, since Ayer

(1936, 108). It is debatable whether the denial should be accepted: Jackson and Pettit (1998) and Suikkanen (2009) argue that the two theory types cannot be meaningfully distinguished, or at least that expressivism has unavoidable 'subjectivist consequences' (Suikkanen 2009, 383). However, I shall assume that we can meaningfully distinguish the two theory types.

According to expressivism, it is no part of ordinary moralising to ask or consider what it is for something to be right or wrong, at least in the sense in which a metaethicist asks this. Instead, the aim of moralising is to find the best way to coordinate our judgements about what to value and do (e.g. Blackburn 1998, 12; Sinclair 2012, 168). By such accounts, the practice of moralising involves no commitment to any (folk) metaethical view; neither realism nor subjectivism, for example.

However, as all metaethicists tend to do, expressivists generally assume that we typically talk *as if* we were moral realists, by consistently talking of moral 'beliefs', 'facts', 'truths', 'objectivity' and so on. They therefore seek to vindicate moral language, so understood. Blackburn's (1998, 79) vindicatory strategy is a quasi-realist one, which relies on his adherence to the deflationary theory of truth. By this theory, any proposition '*p*' is true if and only if *p*, where this encapsulates all that there is to be said about truth. This is then assumed to further deflate talk of facts, properties, beliefs and so on, given that a fact or property is something that truly obtains, that a belief is a truth-apt mental state, and so on. It roughly follows that, when we express an attitude of disapproval towards torture by asserting that 'torture is wrong', we are fully entitled to understand ourselves as expressing a belief that, as a matter of fact, torture is truly wrong.

The many complexities of realist-seeming moral language are then taken to be explicable via the underlying complexities of our moral attitudes, and the task of co-ordinating these. Talking of moral truth, knowledge and belief can help us make sense of, and communicate, our various levels of strength, doubt, or confusion regarding our attitudes. If

we are unsure how to feel about eating meat, for example, we may say that we think that eating meat is permissible, but that we don't know whether to believe this. In contrast, given the horrors of torture, we say that we know that, as a clear matter of fact, torture is objectively wrong. For Blackburn, this statement expresses a first-order moral belief, not a metaethical one. That is, it expresses a firm commitment towards the impermissibility of torture: an attitude of disapproval.

Zangwill (1990) argues that a suitably complex subjectivism might successfully adopt much or all of the quasi-realist strategy, and so similarly vindicate moral language and its features. Roughly in this vein, I argue that emotive subjectivism can be understood, at least to some extent, as a 'modest' version of Blackburn's 'quasi-realism', because it entails that we often usefully talk *as if* there are stance-independent moral facts, truths, and so on (Miller 2005, 77). Unlike Blackburn's 'ambitious' quasi-realist view, however, the modest version does not allow that there *are* stance-independent moral facts, truths and so on. I will, however, argue that we should allow for moral *objectivity*, at least once we understand 'objectivity' in an appropriate way.

The crucial difference between my view and Blackburn's, of course, is that I argue that an utterance like MU, as discussed in Chapter 9, does not express one's moral judgement, but instead refers to it. There is an important potential objection to this view, which I will address now.

### 10.2. The incompatibility objection

The incompatibility objection stems from the simple but obvious fact that we do not appear to be referring to our mental states when we assert moral claims. Consider moral disagreement, as where Steve says that legalising abortion is morally wrong, while Hasina claims that legalising abortion is morally obligatory. According to emotive subjectivism, the *meanings* of

Steve and Hasina's utterances are as follows: Steve means that he disapproves of legalising abortion, whereas Hasina means that she approves of legalising abortion and disapproves of not doing so. Clearly, by this interpretation, neither Steve nor Hasina need be saying anything false. Yet, when Steve and Hasina disagree about whether legalising abortion is morally wrong or obligatory, it seems that at least one of their judgements must be false.

This 'argument from disagreement' certainly *seems* decisive, and it has persuaded many since Moore (2005, 45-46) discussed it in 1912. It seems obvious to us that utterances like MU refer to the objects of our disapproval, and to the properties of these objects, but not to the disapproval itself. Steve and Hasina are not disagreeing about whether they *do* disapprove of legalising abortion, but about whether they *should* do so. Indeed, the relevant kind of disagreement seems to be 'disagreement in attitude' – disagreement about what to value and how to behave – rather than disagreement about facts of any kind (Stevenson 1963, 1).

A similar argument ultimately persuaded Stevenson to disavow any subjectivist elements of his emotivism. Stevenson (1963, 213) came to deny that a sentence like 'Jones ought not to have insulted Smith' includes in its meaning 'I [the speaker] disapprove of Jones' having insulted Smith', because this analysis would lead him to endorse two claims that he rejects. The first is that the former sentence expresses a belief about the speaker's attitude as well as the attitude itself. This is not problematic in and of itself, but Stevenson takes it to imply the second claim: that the speaker's *reasons* for her attitude would include reasons to support her belief that she has the attitude. Stevenson thinks that this is clearly mistaken.

The worry here lies in the fact that, in saying that Jones did something wrong, I am saying something that my interlocuter will agree with only if *she* thinks that Jones did something wrong. If I try to support my claim with reasons to believe that I disapprove, then

even if she accepts these, they need not give her any reason to disapprove. If I am to hope to persuade my interlocuter that Jones did something wrong, then my reasoning *must* be about Jones and her actions. As with the argument from disagreement, this seems to require that MU refers beyond our moral feelings or attitudes. Moral agreement will occur if and only if we both share at least roughly the same attitudes towards Jones and her actions. It will not occur just because we share the same beliefs about our own or each other's moral judgements.

As an objection to emotive subjectivism, these worries can, I think, be generalised to the seeming incompatibility between the claim that MU means, roughly, 'I disapprove of *x*' and the clearly true claims that we do not appear to talk as if we are referring to our own judgements when we moralise, and that we appear to talk instead as if we are referring to the objects of our moral judgements. This kind of general objection is, I take it, at the core of the expressivist's – and, indeed, many people's – denial of simple subjectivism. This is the incompatibility objection.

To overcome the incompatibility objection, we have to explain moral disagreement as *normative* disagreement. Disagreement about whether *x* is wrong is not simply disagreement about how we feel about *x*, but neither is it simply disagreement about the features of *x*. Even if you persuade someone that an action is to be called 'wrong', you are not in agreement unless she has a disapproving and aversive attitude towards it (Stevenson 1963, 16). Disagreement about whether *x* is wrong is disagreement about, roughly, how we should feel and behave towards *x*. Expressivism appears to capture this very well, because it explains moral discourse generally as discourse about what to value or do. Indeed, I believe that it *does* explain this well, and I will argue that emotive subjectivism can borrow much of its explanation, albeit with a focus on the pragmatics of moral language rather than on the meanings of moral terms.

Given emotive subjectivism, where Steve says that legalising abortion is wrong and Hasina says it is right, this does not, of itself, constitute any factual disagreement. Each knows the other's moral judgement (Steve knows that Hasina strongly approves, and Hasina knows that Steve strongly disapproves). Each (let us say) agrees on all other relevant facts. Yet Hasina will not have won the debate until Steve approves of legalising abortion, and vice versa. The kinds of reasons that each needs to persuade the other are reasons about abortion, its consequences, women's preferences, the law, and so on; not about their own psychologies. It is *because* Hasina wants Steve to recognise and respond to the many harms of forbidding safe abortions that she will list these, although in doing so she will also explain much of that which makes her disapprove of torture.

As Haidt (2001; 2012) urges, in presenting her reasons, Hasina will be engaged to some extent in post hoc reasoning. This is because her moral judgement consists of an intuition (roughly, that prohibiting abortion is such as to cause harm, on my account) and because the associative thought processes that lead to this intuition are opaque to her. In Haidt's (2001, 818) terms, the moral reasoning involved in justifying her judgement to Steve will be an 'effortful process, engaged in after a moral judgment is made, in which [she] searches for arguments that will support an already-made judgment'. However, this should not be taken to imply that Hasina's reasoning is defective. In honestly entering the disagreement, Hasina is, *inter alia*, opening herself up to reasons to be persuaded, even if she finds it unlikely that any will be found. She is at least willing to consider new ways of conceptualising abortion, as discussed in §8.4. She is allowing that these may make her change her judgement and behaviours towards abortion, even if she hopes to persuade Steve before she is persuaded. She is thus sincerely engaged in a disagreement in attitude with Steve.

Admittedly, it is not immediately obvious that any subjectivist theory can explain moral disagreement as disagreement in attitude in this way. Köhler (2012), for one, argues that it *cannot* do so, but that expressivism can. Köhler suggests, I think plausibly, that subjectivists and expressivists are primarily concerned to explain our moral thoughts, rather than the meanings of moral sentences. At least, I take this to be his meaning when he says that their theories are '*primarily* theories about which mental states constitute moral *judgements*' (Köhler 2012, 74, Köhler's emphasis). Köhler (2012, 74) assumes that both expressivists and subjectivists typically focus on moral *sentences* because 'moral sentences are those sentences normally used to express moral judgements'. If this is how we must understand the relation between moral sentences and moral judgements, then subjectivism can be understood as 'the view that making a moral judgement is having a belief that one has a certain conative attitude' (Köhler 2012, 76). This certainly suggests that subjectivists cannot plausibly explain moral disagreement. *If* one accepts that moral judgements just are those mental states that are expressed by moral utterances or sentences, then one has to conclude that subjectivism cannot allow that moral disagreement typically involves disagreement in moral judgement. Steve and Hasina simply do not disagree in their moral judgements, so understood (see also Sinclair 2020, 33-34).

This is one reason why I carefully distinguished what I call 'moral judgements' from those mental states expressed by moral utterances. The moral aliefs that I call 'moral judgements' are, I argue, central to our moral concerns, and so similarly central to moral disagreement. When we think about torture, we are, hopefully, deeply concerned about the pain that torture causes. Our disapproval of torture, unlike our judgement that we disapprove, is a crucial element of this concern. When we say that 'torture is wrong', we do so primarily because we want others to feel the same way, or to commit to never supporting torture, and so on. We do not, generally, say 'torture is wrong' primarily because we want to inform

people that we disapprove of torture. However, to understand why we engage in moral discussion as we do, we must look at least as much to the pragmatics of moral language as to the literal meanings of our moral utterances.

According to Köhler, expressivism can explain moral disagreement as subjectivism cannot, because it allows that our conative attitudes are expressed via moral sentences, rather than referred to. On Blackburn's (1998, 70) view, to believe that X is good or right is, roughly, to have 'an appropriately favourable valuation of X', whereas to deny that X is good or right is, roughly, to reject such an attitude. Such attitudes may be directed towards other attitudes and feelings, as well as towards motives or actions (Blackburn 1998, 12). If I utter MU, then my 'attitude is put forward as something to be adopted. The action [of speaking] is one of attempting public coordination or sharing of the attitude' (Blackburn 2006, 151).

Whereas quasi-realist expressivism locates the attempt to share an attitude in the meaning of a moral sentence, emotive subjectivism locates it in the pragmatics of a moral utterance. By saying that some action is morally wrong, I mean that there is something about it that I find seriously unpleasant, although I may not be able to say precisely what, and I implicate my desire that you either persuade me otherwise or adopt the same moral attitude. You will very likely understand this: given the norms of moral assertion, why would I mention my moral judgement unless I hoped to persuade you to converge in your moral judgement? However, different types of moral disagreement and discussion call for different kinds of moral assertions. I turn to some of these now.

### 10.3. Ordinary moral language

Just as Gill (2009a) suspected, as we saw in Chapter 9, non-philosophers appear not to be consistent enough in their use of moral language for us to learn the answers to metaethical questions from any close study of such language. Notably, non-philosophers do not

consistently use seemingly 'objective' moral language (Fisher et al. 2017; Goodwin & Darley 2008; 2012; Pölzler & Wright 2019; 2020; Sarkissian et al. 2011; Wright et al. 2013). In these studies, 'objectivity' is typically contrasted with either 'subjectivity' (Goodwin & Darley 2008; Fisher et al. 2017) or 'relativism' (Sarkissian et al. 2011; Wright et al. 2013). It is typically assumed that moral objectivity is closely related to (if not identical to) moral realism (e.g. Goodwin & Darley 2008, 1340).

To test whether people view morality as 'objective' or 'subjective', these studies typically presented people with a range of statements, including moral statements such as 'robbing a bank in order to pay for an expensive holiday is a morally bad action' (Goodwin & Darley 2008, 1361). They then asked various questions about each statement; typically, mainly of two general kinds. The first kind of questions assess whether the respondent thought that the statement was true, false, or 'just an opinion or attitude' (Wright et al. 2013, 339). Other questions, following the first kind of question, aimed to assess whether the respondents thought it would be possible or not for two people to disagree in their answer to the statement, and to both be correct. These question types closely follow Goodwin and Darley's suggestion that 'if an individual takes a particular ethical claim to be true, and regards situations of ethical disagreement as necessarily implying that at least one party is *mistaken*, then they are an objectivist (with respect to that statement)' (Goodwin and Darley 2008, 1341-2, their emphasis). Fisher et al. (2017, 1123), Sarkissian et al. (2011, 484), and Wright et al. (2013, 339) at least broadly follow Goodwin and Darley's approach.

Pölzler and Wright (2020, 55) somewhat similarly 'take the moral realism/anti-realism debate to be about whether moral sentences are true in that they match (objective) moral facts; and about whether these facts are objective in that they would obtain even if observers had different or no mental states towards them'. However, they went to greater lengths than the other study developers to consider a range of metaethical views in at least

some depth. In their studies, respondents were very clearly directed to answer metaethical questions, after having been given very brief explanations of metaethical positions to respond to. Unsurprisingly, they appeared to respond with confusion, and they struggled to make sense of this new way of thinking, as we all do when first introduced to metaethics. In the other studies, however, people were asked to think more about questions of truth and disagreement than about any explicitly defined metaethical theories.

These studies generally showed that people often provide very different answers to one or both of the two kinds of questions, regarding statements which, they often explicitly recognised, were all moral statements. Some moral statements were treated objectively (in the sense here discussed): the respondents answered that a statement was either true or false, and that if two people disagreed about the statement then one of them must be mistaken. Other moral statements were treated non-objectively to some extent, in that it was allowed either that a statement was neither true nor false, or that two people could disagree without either person being mistaken, or both.

The studies suggest that different *types* of moral judgement are treated in different ways. To give some examples: Negative moral judgements are treated as more objective than positive ones; judgements that are more firmly held are treated as more objective than ones that are less firmly held; judgements that cause discomfort when others disagree with them are treated as more objective than ones that cause less or no such discomfort (Goodwin & Darley 2008; 2012). Goodwin and Darley (2008, 1360) also noted, to their 'surprise', that people 'who grounded their ethical beliefs in their pragmatic consequences for society tended to be more objective than those who did not'. People's 'intuitions take a strikingly relativist turn when they are encouraged to consider individuals from radically different cultures or ways of life', or so it appears (Sarkissian et al. 2011, 500). People also appear to have more objectivist intuitions when they think there is a general moral consensus, when they strongly

desire to punish transgressors, and when they feel disgusted by other cultures' practices (Pölzler & Wright 2019, 4). This list is not exhaustive.

Pölzler and Wright (2019; 2020) and Wright et al. (2013) argue from these studies that non-philosophers are metaethical pluralists. However, I think that the studies can be at least as plausibly interpreted to show that people typically have no metaethical commitments or, perhaps even more plausibly, that any metaethical commitments that they might have are significantly less important to them than are their *moral* commitments. It is their moral commitments that people seemed to want to focus on throughout the studies; not their metaethical understanding of these commitments. People were asked, for example, about whether some actions are wrong, whether these actions are truly wrong, and whether other people might be 'mistaken' if they disagree (Goodwin & Darley 2008, 1348). I suggest that (at least) those of us who are not already metaethicists will typically treat all these as moral, rather than metaethical, questions. We might well say that the person who disagrees with the claim that abortion is permissible is 'mistaken', not because of any metaethical view that we hold, but because we think that such people are *morally* mistaken: we disapprove of their attitudes.

There is one possible exception, I think. The more that respondents justify 'morality by reference to God', the more their 'intuitions' suggest that there is some kind of 'moral objectivity' (Pölzler & Wright 2019, 4). It may be that the relevant religious belief systems include metaethical theories, so that religious people *are* discussing their metaethical commitments in such cases. However, I am far from certain of this: they may be stressing their moral commitments in cases where they are unwilling to seriously consider altering these commitments, *because* of their religious beliefs.

In most or all other cases, I suggest that people appear to treat moral statements 'objectively' when they, roughly, want to assert or to stand by these statements rather than to

open themselves up to other viewpoints, and that they treat moral statements 'non-objectively' in other cases. From the perspective of the respondents of these studies, it appears very likely that they were responding to moral questions, such as whether abortion is morally wrong, or whether other people must be morally mistaken if they disagree. It seems very likely that the variations in the kinds of answers that they gave is due to the fact that they have been *moralising*: they have attempted, as best they can, to decide what to do, where to persuade others, and where to open themselves to being persuaded.

A good example to support this explanation comes from Fisher et al.'s (2017) study, in which people tended to treat their moral judgements concerning 'highly controversial topics' as more 'objectively true' when they were trying to win arguments against others than when they were engaged in cooperative interactions. Fisher et al. (2017, 1132) conclude that the study 'demonstrated that the character of people's social interactions influences their understanding of truth'. However, I think it far more likely that these shifts were made for pragmatic reasons than because the people concerned changed their metaethical views – or pretended to – between different interactions. We can, I suggest, make better sense of these shifts by understanding the people involved to be less concerned, if at all, with questions about metaethical truth, even broadly construed, than they were with normative questions: questions of what to value or how to act.

Wright et al. (2013, 354) allow that all such debates are genuinely metaethical, but they nevertheless suggest a similar explanation to this: that the classification of some moral judgements as 'objectively grounded', and others as 'relatively grounded,' may 'serve the important psycho-social function of determining the level of permissible dialogue and exploration'. The *function* of talking in a realist manner seems to be to remove a moral issue from 'the realm of legitimate personal and social negotiation': to demonstrate that one not only wants to abstain from certain actions, but to demand of others that they too abstain; to

signal the desire to punish transgressors; and so on. In contrast, we create 'space for personal and social exploration, discussion, and debate' by treating a morally salient issue in less objective terms (Wright et al. 2013, 355).

Wright et al.'s interpretation is grist to the expressivist's mill. If we are prepared to accept a deflationary theory of truth, then this may well support a quasi-realist view of some kind, albeit one that is presumably adjusted to account for people's inconsistent approach to moral language. However, I think it also supports my theory, which requires no adherence to any one theory of truth, perhaps even more than it supports expressivism.

Recall, from §9.1, the claim that the meaning of moral sentences can be either *strongly* or *weakly* indexical. Where a sentence is used in a strongly indexical sense, then we mean that we disapprove of *x*, here and now. Where a sentence is used in a weakly indexical sense, then the meaning of the sentence is more closely related to the tense of the sentence.

I suggest that Wright et al.'s 'relatively grounded' moral utterances are weakly indexical ones, which are employed more to reach agreement than to demonstrate commitment to one's moral judgement. If we contemplate practices in different cultures, for example, then we may feel that we can potentially learn from these practices, or perhaps that it matters less to us if people act in ways that we find immoral in such cultural contexts than if they perform similar actions within our own cultural contexts. Perhaps even more plausibly, we might consider it unlikely that we could (or should) successfully *demand* of individuals in different cultural contexts that they change their practices, although we might be able to persuade them via more open means of discussion, negotiation, and so on.

In cases like this, we would do well to acknowledge the feelings and judgements of others, and to reflect on how we would feel in such contexts. Weakly indexical moral sentences are useful in such contexts. To say that, 'if I were a member of your culture, then it would not be wrong for me to act as you do', when said in a weakly indexical sense, is to

acknowledge that if I *had* been brought up in your culture, then I would not disapprove of so acting. Such sentences may reflect an openness to learning, or an acknowledgement that one is engaged in a debate with moral equals.

'Objectively grounded' utterances are, I suggest, strongly indexical ones, used more to demonstrate commitment than to reach agreement. If a moral commitment is important to us, or if we feel a strong desire to punish transgressors, or if we think that there is a general moral consensus, then we may well be less inclined to open ourselves up to discussion about the commitment than we would be otherwise. This may be because we feel less comfortable about doing so (another predictor of moral objectivity), or because we feel confident that such discussions have already been satisfactorily concluded. Similarly, if we believe that there are pragmatic consequences when people do not behave in accordance with our commitments, or if we feel disgusted by those who behave otherwise, then the commitments are likely to be ones that we will not want to seriously consider altering.

In cases like this, we are very likely to stand firm, to remain heavily influenced by our current moral feelings, and to refuse to seriously consider whether we would feel differently in other contexts. Strongly indexical moral sentences are useful in such contexts. In such sentences, to say that something was, would, or will be wrong is to say that it was, would, or will be such that, thinking about it now, I disapprove of it. For example, to say that, 'if I were in your shoes, then I might approve of your action, but it would still be wrong', when said in a strongly indexical sense, is to say that I might, in other circumstances, approve of the action, but it would still be such that I disapprove of it *here and now*. Such sentences demonstrate the strength of one's commitment, and of one's determination to adhere to it.

I believe that the notion of strong indexicality and the opacity thesis are sufficient to allow that we can, sincerely and truly, say things like 'if I were in your shoes, then I might approve of your action, but it would still be wrong', or 'if I were to stop disapproving of

torture, then torture would still be wrong'. According to emotive subjectivism, when the latter sentence, for example, is uttered in a *weakly* indexical sense, it means that, even if I were to stop disapproving of torture, then I would still disapprove of torture. This expresses a clear contradiction. However, when the same sentence is uttered in a *strongly* indexical sense (as I believe that such sentences typically are), it means, 'if I were to stop disapproving of torture, then the torture of which the hypothetical version of myself no longer disapproves would still be something towards which I feel disapproval, as I think about it *here and now*'. No contradiction is involved.

The opacity thesis is crucial here, in understanding why we utter sentences with this meaning, and why they do not appear to us to mean this when we do utter them. Most of what we think when we judge something wrong is unavailable to conscious attention. We usually recognise *when* we judge something wrong and, as discussed in Chapters 7 and 8, we can typically clearly distinguish these kinds of intuitive responses from other kinds, but we cannot know, purely from introspection, what it is for something to be wrong. Moreover, I have argued, it is not greatly important to most of us, most of the time, what it is for something to be wrong, where this is considered as a metaethical question.

According to emotive subjectivism, we often shift between using weakly and strongly indexical senses of moral sentences, and we do so largely for pragmatic reasons. To argue for this view in greater detail, I will aim to show that several of Blackburn's quasi-realist strategies for vindicating moral language can be applied to emotive subjectivism. In §10.4, I at least begin to argue that the emotive subjectivist can vindicate moral language as well as the quasi-realist can – perhaps better – and that this is also an improvement on some other metaethical theories that endorse the Humean ontology.

## 10.4. Towards a vindication of moral language

I cannot hope to provide a full account of the ways that emotive subjectivism might vindicate moral language. Instead, I will try to consider some of the most significant aspects of moral language that might seem hard to vindicate by this theory, and to demonstrate that there appear to be very similar strategies for the emotive subjectivist to employ to those developed within Blackburn's quasi-realism.

### *10.4.1 Moral objectivity*

At least sometimes, our moral language appears to suggest that we are referring to stance-independent moral facts, as discussed at the beginning of this chapter. Blackburn provides a compelling argument that it is not only realists who can explain the existence of such facts, but that expressivists can too:

> [Realists say that denying] women the vote is wrong, whatever your [sic] or I or anyone else thinks. Can an expressivist say as much? This is to be assessed in the standard way, of imagining scenarios or possible worlds in which you or I or others think that women should not have a vote, and passing a verdict on them. Naturally, these scenarios or possibilities excite condemnation, and so the answer is that denying women the vote is wrong, whatever you or I or anyone else thinks about it. In giving that answer one is, of course, standing within one's own moral view. One is assessing the scenario in the light of things one thinks and feels about such matters. But that is no objection, since there is no other mode of assessment possible. One cannot pass a verdict without using those parts of one's mind that enable one to pass a verdict (Blackburn 2006, 154).

We have seen that, given Blackburn's deflationary theory of truth, he can move quickly from the argument above to the further conclusion that the relevant 'verdict' – that denying women the vote is wrong, whatever anyone thinks – expresses a true belief about a moral fact. However, we also saw in §10.3 that moralising does not always, or even typically, require or include reference to such facts: people are not implicit realists. We need not explain how anything like stance-independent moral facts or properties might truly exist, for such claims are not sufficiently widespread in ordinary moral language to require vindication. We must, however, explain why we often, but not always, treat our moral judgements as 'objective', as the term is used in §10.3.

I think the emotive subjectivist can use a similar argument to Blackburn's to explain this. If I think of a scenario or possible world in which women are denied the vote and in which I fail to disapprove of women being denied the vote, then I experience disapproval. I disapprove of my own attitude in this scenario, and of any similar attitude, and so I call all such attitudes 'wrong'. Here and now, standing within my 'own moral view', to use Blackburn's term, I deem it wrong to deny women the vote, no matter what anyone, including hypothetical versions of myself, might think.

By emotive subjectivism, anything of which I disapprove is, by definition, really and truly wrong (I really do, as a matter of fact, disapprove of it). I can truthfully *say*, therefore, of anything that I judge wrong, 'it *really* is wrong!' Here, the stress on 'really' is simply a useful way to propagate my attitude.

Sinclair (2014, 430-431) suggests that an expressivist account of moral disagreement can profit from seeing a parallel with 'political disputes', broadly conceived. Just as two communities can stand firm while remaining open to negotiations when negotiating some contested territory, so two people may stand firm but remain open to negotiations during

moral disagreements. This is a helpful analogy, from which emotive subjectivists may similarly benefit. In her debate with Steve, Hasina is strongly committed to her moral judgement, and she has important and reflectively endorsed reasons for this. However, she also recognises that the question of the permissibility or otherwise of abortion is one that ultimately requires a communal response. Because she cares deeply about the permissibility of abortion, she may therefore concentrate her moral language on what is, when considered merely as a matter of meaning, the trivially true claim that banning abortion is, *as a matter of fact*, wrong. She may say that it is 'universally' wrong, by which she will mean that she currently feels disapproval towards the idea of anyone, anywhere, being denied access to safe abortion. Much as discussed in §8.3, she is likely to support such claims with non-moral claims about the harms caused by denying abortion, the freedom that the legal right to abortion gives women, and so on.

As Stevenson (1963, 1) saw, the kind of moral disagreement that exists in cases like Steve and Hasina's closely parallels the kinds of factual disagreement in which we are unwilling to back down. Hasina treats her disapproval of banning abortion as a fixed point, as she would a well-evidenced belief that legalising abortion decreases infant mortality, for example. In either kind of case, she can best proceed by presenting reasons for her attitude, while assuming that any reasonable opponent should be compelled by her arguments to either share her attitude or offer strong arguments of his own. Therefore, she is well served by talking *as if* engaged in factual disagreement. 'Either it's wrong or it's not', she might say. 'Can't you see that *these* features are the morally relevant ones?' As Stevenson saw, this kind of language facilitates the practical, coordinating role of morality.

Generally, where we appear to talk as if discussing a stance-independent moral reality, we are being objective, in that we are using strongly indexical moral language. If I consider a misogynist's claim that only men should vote, then I experience disapproval, and I

sincerely and accurately call the claim 'wrong'. Given my metaethics, I must allow that his misogyny is only wrong in virtue of my disapproval of it, and of course I allow that he approves of it. However, I need not conclude that his misogyny is morally right, even relative to his outlook. As Blackburn (1998, 314) argues, the *descriptive* claim that others approve of things of which we disapprove is very different from the *moral* claim that theirs is 'no worse a view' than ours. So, the misogynist approves of his misogyny. If *I* consider whether misogyny is therefore morally permissible in any way, then I must moralise, and if I moralise then I must call misogyny 'wrong'. This is a commitment which I am not willing to consider altering, or at least, not just because a misogynist disagrees with it. I therefore remain objective: I say, 'no matter what anyone thinks, it really is wrong to deny women the vote', and I mean this in a strongly indexical sense. Denying women the vote would be morally right only if I were to approve, here and now.

Before considering other forms of moral language, I will briefly argue that emotive subjectivism is better placed to vindicate moral objectivity, as I understand it, than any dispositional theory of value.

### *10.4.2. Dispositional theories of values*

The most influential dispositional theory of value is Lewis's (1989, 113), which roughly holds that something is a value 'if and only if we would be disposed, under ideal conditions, to value it'.[1] To be 'a value – to be good, near enough – means to be that which we are disposed, under ideal conditions, to desire to desire' (Lewis 1989, 116). And ideal conditions are of those 'of the fullest possible imaginative acquaintance' (Lewis 1989, 121). Therefore,

[1] The full definition is: 'Something of the appropriate category is a value if and only if we would be disposed, under ideal conditions, to value it' (Lewis 1989, 113). Here, we need not worry about the 'appropriate category'.

if we want to know whether actions, motives or other things are good or not, then we have to approximate this level of imaginative acquaintance as best we can, by imagining 'vividly and thoroughly how it would be if these putative values were realised (and perhaps also how it would be if they were not)' (Lewis 1989, 121). If we desire to desire something after subjecting it to a rigorous process of imaginative acquaintance, then we can be at least fairly confident that it is good.

As Lewis (1989, 123) stresses, this allows for cases of knowledge and cases of 'ignorance and error, for hesitant opinion and modesty, for trying to learn more and hoping to succeed'. It allows for moral improvement to occur if we think carefully about the objects of our attitudes, and it allows that we may learn from the experiences of others. These are all undoubtedly positive features. Nevertheless, as Lewis concedes, it is highly possible that we will not all be disposed to value the same things, even in ideal circumstances.

If this possibility obtains, then we cannot simply see all values as values *simpliciter*. Giving the vote to women might be good according to our value system, but it is bad by the misogynist's. If female emancipation is not something that the misogynist would value under ideal conditions, then I must simply accept that it *is* bad for him. However, as Egan (2012, 565) argues, when I think this, my thought of badness will not be an evaluative thought, for I do not share the relevant value system. Prinz (2007) accepts this kind of view, as does Dreier (1990; 2009) in his speaker relativism. However, I agree with Egan (2012, 565) that it is an 'uncomfortable and unsatisfying result' of dispositionalism. We want to deny that misogyny is in *any* way good or valuable, without any error or omission, and without requiring anything more than the Humean ontology.

Egan (2012, 559) argues that we can do so, by developing a '*de se* relativist version of a dispositional theory of value'. The idea is that we can adopt, for example, something very like Lewis's account, but where 'evaluative belief is not a matter of believing that *x* is

disposed to cause [response] R in [person(s)] Ks in [conditions] C, but of self-attributing the property, *being someone in whom x is disposed to cause R in C*.' As Egan (2012, 573) concedes, this is primarily an account of evaluative belief rather than language, and so he requires a 'broadly Stalnakerian' theory of the semantic content of sentences. Roughly, an utterance of MU is, by convention, understood by its hearers as a 'bid' to add the property *being someone in whom x is disposed to cause disapproval in C* 'to the stock of things that all of the parties to the conversation believe, take each other to believe, etc.' (Egan 2012, 573).

I suggest that emotive subjectivism can allow for something very similar, but without requiring anything like Stalnaker's theory of meaning. Moreover, although Egan's suggested form of dispositionalism responds to similar worries to my own, I think we do better without dispositionalism. Where we call some distant act of unfairness 'wrong', we do so, not primarily because we care about how (we believe) it would feel to us, or to anyone, under other conditions, but simply because we object to unfairness, here and now.

Emotive subjectivism can, I believe, explain our tendencies towards moral objectivity, including our rejection of moral relativism – where we do reject this – better than typical dispositional theories can, while remaining relatively neutral on theories of truth and semantic meaning. Yet on other occasions, a different kind of language is needed.

### *10.4.3. Moral subjectivity*

If you tell me, as a decreasingly frequent meat-eater, that eating meat is impermissible, then I will be open to learning something. I will allow that different people have good reasons for holding different attitudes, and I will want to hear yours. Here, I am less interested in persuading you than in discovering whether you might persuade me. This case is less similar to a factual argument than to ones where I might be persuaded to persist with trying to enjoy some new flavour, for example, or to devote some time to a musician whose music has not

yet appealed to me. Here, I *do* seem to talk as simple subjectivists predict: 'I don't feel that way about it yet, but I suspect I might do so at some point'.

In Chapter 8, I discussed various ways in which we can put pressure on one another to rethink the *types* of actions or motives that we are morally assessing, so that we may potentially feel different moral aliefs toward them. I suggest that, where we are open to reconceptualising the objects of our moral judgements and rethinking the judgements themselves, we signal this by talking about these judgements in less objectivist language than otherwise: we use MU in a weakly indexical sense. This allows 'space for personal and social exploration, discussion, and debate', as Wright et al. (2013, 354) suggest. In this kind of case, I will talk less of moral truths and more about different moral views. 'I am coming to think it could be wrong', I might say, 'although I don't disapprove myself'. By this, I mean that I am coming to think that eating meat is possessed of features of kinds that I elsewhere disapprove of, or otherwise might disapprove of.

This type of language is used wherever we seriously consider our own moral doubt or error. Any plausible metaethical theory must allow that we can experience disapproval without being certain that what we disapprove of is wrong. As Blackburn (1998, 318) does within his quasi-realism, the emotive subjectivist must explain and vindicate our talk of fears for our own fallibility.

Blackburn (1998, 318) argues that, to doubt one's own moral beliefs is to consider first-order evaluative questions about whether they stand up to the kinds of moral thinking in which one ought to engage. It is to accept that, if one improved one's thinking about the case, one might change one's mind. If I am unsure whether I am correct in judging it permissible to eat meat, then I am acknowledging that I might come to judge otherwise, perhaps if I were to learn more information or think more sensitively about the pain of animals.

I think that the emotive subjectivist can, once more, adopt this response, or one very like it, by reference to weakly indexical moral utterances and the fact that we may feel very differently about action types or tokens depending on how we conceptually categorise them. This is surely something that is readily apparent to us. Indeed, consider Blackburn's (2006, 155) very brief example of a case of moral doubt, as follows: 'Was he selfish and despicable, or prudently protecting himself in an unfortunate situation?' Similarly, the meat-eater might ask, 'is it really cruel, or is it humane? Are we killing beings with hopes and dreams, or merely animals that cannot be capable of such things?' Different ways of thinking about the matter will elicit different feelings. Some feelings may clash with others, and some may not seem to relate to our typical justificatory reasons for holding moral judgements. In trying to make sense of this, we try to come to *stable* moral judgements.

Here, I follow the quasi-realist line, by which a 'stable' moral judgement is one that would survive any of the kinds of reflection of which the judger approves, or which she would approve of on further reflection (Egan 2007). For the quasi-realist, any question about whether one's own moral belief is true is a question about whether it is stable. The emotive subjectivist can follow a very similar path. It may be true that meat-eating is permissible insofar as I think of it as such, but *simultaneously* true that it is wrong insofar as I think of it as eating animals that people have killed. If I am to reach a verdict that it is simply one or the other, then I must reach a stable moral judgement regarding meat-eating.

In such cases, I may employ weakly indexical thoughts or utterances to consider what kinds of features are such as to make me approve or disapprove, and whether I approve or disapprove of my responses towards such features. 'I might come to disapprove if I were faced with the pains of animals', I might say. Where I am relatively free of doubt, I will be more likely to employ strongly indexical thoughts or utterances, although I may have other reasons too for changing between these.

Sinclair (2012, 174) claims that '[t]hree important assumptions of moral practice are that moral disagreements are possible, that moral discussion is sometimes a fruitful way of resolving such disagreement and that reasons can be offered for and against moral claims'. I hope to have shown that emotive subjectivism has the resources to explain how and why these assumptions are correct. Of course, much more remains to be said.

To hint at the answer to one further question, emotive subjectivism can, I think, allow for moral progress, as well as for moral doubt. We typically approve of consistent moral attitudes, and so we can and, plausibly *do,* each try to develop a consistent and coherent set of stable moral judgements. As Lenman (1999, 167) argues, the absence of anything like a rational justification for our fundamental moral desires nevertheless allows that '[w]ithin the system of values, interests and institutions we inhabit there is plenty for the justification of ethical and other claims to be'. We can aim to improve and justify our moral judgements, by reference to the kinds of thinking, valuing and feeling of which we approve.

This may not be easy to achieve. It may not even be possible, or not for everyone, for we cannot fundamentally change many of our moral aliefs. However, we can rethink how to conceptualise and categorise action and motive types, and we can thereby influence how we feel about them. If it is habit that causes us to form our moral judgements, then – as Aristotle argued long before Hume – we should try to get into good habits. And given the fundamental relationship between harm and judgements of wrongness, these should be primarily habits of compassion and care for others.

I will conclude with a further quotation from Lenman: one that is particularly salient for emotive subjectivism, given this theory's thesis of strong indexicality as well as its focus on the habitual and social causes of our moral judgements:

[T]he fear that we might… become, by our present lights, contemptible, that the moral perspective from which our present attitudes make sense might be lost to us, is not always idle fantasy but may signal real dangers… The disintegration of our communities of judgement is a real danger but faith in a spurious objectivity for our values cannot meet it… If it is [just politics and education] that shores up our values…, the proper moral is not that we may no longer take our values seriously but rather that politics and education should be taken very seriously indeed (Lenman 1999, 174).

# Bibliography


Abramson, Kate. (1999). Correcting *Our* Sentiments about Hume's Moral Point of View. *The Southern Journal of Philosophy* 37(3): 333-361.

———. (2001). Sympathy and the Project of Hume's Second Enquiry. *Archiv für Geschichte der Philosophie* 83(1): 45-80.

Aiken, Henry. (1979). An Interpretation of Hume's Theory of the Place of Reason in Ethics and Politics. *Ethics* 90(1): 66-80.

Alanen, Lilli. (2006). Powers and Mechanisms of the Passions. In *The Blackwell Guide to Hume's Treatise*, ed. Saul Traiger, 179-198. Malden, MA: Blackwell Publishing Ltd.

Alvarez, Maria. (2010). *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford: Oxford University Press.

Árdal, Páll S. (1966). *Passion and Value in Hume's Treatise*. Edinburgh: Edinburgh University Press.

———. (1977). Another Look at Hume's Account of Moral Evaluation. *Journal of the History of Philosophy* 15(4): 405-421.

———. (1989). Hume and Davidson on Pride. *Hume Studies* 15(2): 387-394.

Ayer, A.J. (1936). *Language, Truth and Logic*. London: Penguin.

———. (1980). *Hume: A Very Short Introduction*. Oxford: Oxford University Press.

Baier, Annette. (1991). *A Progress of Sentiments: Reflections on Hume's Treatise*. Cambridge MA: Harvard University Press.

Barker, Stephen J. (2000). Is value content a component of conventional implicature? *Analysis* 60(3): 268–279.

Baron, Marcia. (2001). Hume's Noble Lie. In *Hume: Moral and Political Philosophy*, ed. Rachel Cohon, 273–290. Aldershot/Burlington: Ashgate/Dartmouth.

Beebee, Helen. (2006). *Hume on Causation*. Oxon: Routledge.

Besser-Jones, Lorraine. (2006). The Role of Justice in Hume's Theory of Psychological Development. *Hume Studies* 32(2): 253–276.

Blackburn, Simon. (1984). *Spreading the Word: Groundings in the Philosophy of Language*. Oxford: Oxford University Press.

———. (1993a). *Essays in Quasi-Realism*. Oxford: Oxford University Press.

———. (1993b). Hume on the mezzanine level. *Hume Studies* 19(2): 273-288.

———. (1998). *Ruling Passions*. Oxford: Oxford University Press.

———. (2006). Antirealist Expressivism and Quasi-Realism. In *The Oxford Handbook of Ethical Theory*, ed. David Copp, 146-162. Oxford: Oxford University Press.

Boehm, Miren. (2013). The Normativity of Experience and Causal Belief in Hume's *Treatise*. *Hume Studies* 39(2): 203-231.

Boisvert, Daniel R. (2008). Expressive-Assertivism. *Pacific Philosophical Quarterly* 89(2): 169–203.

Boyd, Richard. (1988). How to be a Moral Realist. In *Essays on Moral Realism*, ed. Geoffrey Sayre-McCord, 181-228. Ithaca, NY: Cornell University Press.

Bricke, John. (2000). *Mind and Morality: An Examination of Hume's Moral Psychology*. Oxford: Oxford University Press.

Brink, David O. (1986). Externalist Moral Realism. *Southern Journal of Philosophy* 24(S1): 23-41.

———. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge: Cambridge University Press.

Brown, Charlotte. (1994). From Spectator to Agent: Hume's Theory of Obligation. *Hume Studies* 20(1): 19-36.

———. (2001). Is the General Point of View the Moral Point of View? *Philosophy and*

*Phenomenological Research* 62(1): 197-203.

Capaldi, Nicholas. (1989). *Hume's Place in Moral Philosophy*. New York: Peter Lang Publishing Inc.

Carlson, Åsa. (2014) The Moral Sentiments in Hume's *Treatise*: A Classificatory Problem. *Hume Studies* 40(1): 73-94.

Cohon, Rachel. (1997). Is Hume a Noncognitivist in the Motivation Argument? *Philosophical Studies* 85(2/3): 251-266.

———. (2008). Hume's Indirect Passions. In *A Companion to Hume*, ed. Elizabeth S. Radcliffe, 159-184. Malden, MA: Blackwell Publishing Ltd.

———. (2010). *Hume's Morality: Feeling and Fabrication*. Oxford: Oxford University Press.

Cohon, Rachel and Owen, David. (1997). Hume on Representation, Reason and Motivation. *Manuscrito* 20: 47-76.

Copp, David. (2001). Realist-Expressivism: A Neglected Option for Moral Realism. *Social Philosophy and Policy* 18(2): 1-43.

———. (2009). Realist-Expressivism and Conventional Implicature. In *Oxford Studies in Metaethics*, vol. 4, ed. Russ Shafer-Landau, 167-202. Oxford: Oxford University Press.

Cunningham, William. A., Carol. L. Raye and Marcia. K. Johnson. (2004). Implicit and explicit evaluation: fMRI correlates of valence, emotional intensity, and control in the processing of attitudes. *Journal of Cognitive Neuroscience* 16(10): 1717–1729.

Darwall, Stephen. (1993). Motive and Obligation in Hume's Ethics. *Noûs* 27(4): 415-448.

———. (1994). Hume and the Invention of Utilitarianism. In *Hume and Hume's Connexions*, ed. M. A. Stewart and John P. Wright, 58-82. Edinburgh: Edinburgh University Press.

———. (1997). Hutcheson on Practical Reason. *Hume Studies* 23(1): 73-90.

Dauer, Francis W. (1999) Force and Vivacity in the *Treatise* and the *Enquiry*. *Hume Studies* 25(1/2): 83-100.

Davidson, Donald. (1976). Hume's Cognitive Theory of Pride. *The Journal of Philosophy* 73(19): 744-757.

Davie, William. (1998). Hume's General Point of View. *Hume Studies* 24(2): 275-294.

Davis, Wayne A. (1998). *Implicature: Intention, Convention, and Principle in the Failure of Gricean Theory.* Cambridge: Cambridge University Press.

———. (2003). *Meaning, Expression, and Thought.* Cambridge: Cambridge University Press.

De Villiers-Botha, Tanya. (2020) Haidt et al.'s case for moral pluralism revisited. *Philosophical Psychology* 33(2): 244-261.

DeScioli, Peter, Sarah S. Gilbert and Robert Kurzban. (2012). Indelible victims and persistent punishers in moral cognition. *Psychological Inquiry* 23(2): 143-149.

Debes, Remy. (2007a). Humanity, Sympathy and the Puzzle of Hume's Second *Enquiry*. *British Journal for the History of Philosophy* 15(1): 27-57.

———. (2007b). Has anything changed? Hume's theory of association and sympathy after the *Treatise*. *British Journal for the History of Philosophy* 15(2): 313-338.

Dreier, James. (1990). Internalism and Speaker Relativism. *Ethics* 101(1): 6-26.

———. (2009). Relativism (and Expressivism) and the Problem of Disagreement. *Philosophical Perspectives* 23(1): 79–110.

Egan, Andy. (2007). Quasi-realism and fundamental moral error. *Australasian Journal of Philosophy* 85(2): 205-219.

———. (2012). Relativist Dispositional Theories of Value. *The Southern Journal of Philosophy* 50(4): 557-582.

Evans, Jonathan St. B. T. and Keith E. Stanovich. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* 8(3): 223-241.

Fieser, James. (1992). Hume's Classification of the Passions and Its Precursors. *Hume Studies* 18(1): 1–17.

Finlay, Stephen. (2014). *Confusion of Tongues: A Theory of Normative Language*. Oxford: Oxford University Press.

Fisher, Matthew, Joshua Knobe, Brent Strickland, Frank C. Keil. (2017). The Influence of Social Interaction on Intuitions of Objectivity and Subjectivity. *Cognitive Science* 41: 1119–1134.

Fletcher, Guy. (2014). Moral Utterances, Attitude Expression, and Implicature. In *Having it Both Ways*, ed. Guy Fletcher and Michael Ridge, 173-198. Oxford: Oxford University Press.

Fodor, Jerry Alan. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Flew, Antony. (1963). On the Interpretation of Hume. *Philosophy* 38(144): 178-182.

Garrett, Don. (2002). *Cognition and Commitment in Hume's Philosophy*. Oxford: Oxford University Press.

———. (2006). Hume's naturalistic theory of representation. *Synthese* 152: 301-319.

———. (2007). The First Motive to Justice: Hume's Circle Argument Squared. *Hume Studies* 33(2): 257–288.

———. (2015). *Hume*. Oxon: Routledge.

Gauthier, David. (1992). Artificial Virtues and the Sensible Knave. *Hume Studies* 18(2): 401-428.

Gawronski, Bertram and Galen V. Bodenhausen. (2006). Associative and Propositional Processes in Evaluation: An Integrative Review of Implicit and Explicit Attitude

Change. *Psychological Bulletin* 132(5): 692-731.

Gendler, Tamar Szabó. (2008a). Alief and belief. *The Journal of Philosophy* 105(10): 634-663.

———. (2008b) Alief in Action (and Reaction). *Mind and Language* 23(5): 552-585.

———. (2011). On the Epistemic Costs of Implicit Bias. *Philosophical Studies* 156(1): 33-63.

Gibbard, Allan. (2003). *Thinking How to Live.* Cambridge, MA: Harvard University Press.

Gill, Michael B. (1996). Fantastick Associations and Addictive General Rules: A Fundamental Difference between Hutcheson and Hume. *Hume Studies* 22(1): 23-48.

———. (2009a). Indeterminacy and variability in meta-ethics. *Philosophical Studies* 145(2): 215-234.

———. (2009b). Moral Phenomenology in Hutcheson and Hume. *Journal of the History of Philosophy* 47(4): 569-594.

———. (2010). *The British Moralists on Human Nature and the Birth of Secular Ethics*. Cambridge: Cambridge University Press.

Goodwin, Geoffrey P., and John M. Darley. (2008) The psychology of meta-ethics: Exploring objectivism. *Cognition* 106(3): 1339-1366.

———. (2012) Why are some moral beliefs perceived to be more objective than others? *Journal of Experimental Social Psychology* 48(1): 250-256.

Graham, Jesse, Haidt, Jonathan and Brian A. Nosek. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*. 96(5): 1029-1046.

Graham, Jesse, Brian A. Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva and Peter H. Ditto. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*. 101(2): 366-385.

Gray, Kurt, Chelsea Schein and Adrian. F. Ward. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*. 143(4): 1600–1615.

Greenwald, Anthony. G., Debbie McGhee and Jordan Schwartz. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology* 74(6), 1464–1480.

Greig, J.Y.T. (2011). *The Letters of David Hume Volume 1: 1727 – 1765.* Oxford: Oxford University Press (HL1).

Grice, H.P. (1968). Utterer's Meaning, Sentence-Meaning, and Word-Meaning. *Foundations of Languag*e. 4(3): 225-242.

———. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Haidt, Jonathan. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review.* 108(4): 814-834.

———. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. London: Allen Lane.

Haidt, Jonathan, and Fredrik Bjorklund. (2008). Social Intuitionists Answer Six Questions about Moral Psychology. In *Moral Psychology Volume 2: The Cognitive Science of Morality: Intuition and Diversity*, ed. Walter Sinnott-Armstrong, 181-217. Cambridge MA: MIT Press.

Haidt, Jonathan and Jesse Graham. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*. 20(1): 98–116.

Haidt, Jonathan, Jesse Graham and Peter H. Ditto. (2015). A straw man can never beat a shapeshifter: Response to Schein and Gray (2015). [Online] Available at http://www.yourmorals.org/blog/2015/10/a-straw-man-can -never-beat-a-

shapeshifter/

Haidt, Jonathan and Craig Joseph. (2008). The Moral Mind. In *The Innate Mind, Volume 3: Foundations and the Future*, eds. Peter Carruthers, Stephen Laurence and Stephen Stich, 367-392. Oxford: Oxford University Press.

Harris, James A. (2010). Hume on the Moral Obligation to Justice. *Hume Studies* 36(1): 25–50.

———. (2015). *Hume: An Intellectual Biography*. New York: Cambridge University Press.

Harrison, Jonathan. (1976). *Hume's Moral Epistemology*. London: Oxford University Press.

———. (1981). *Hume's Theory of Justice.* Oxford: Clarendon Press.

Hearn, Thomas K. (1976). General Rules and the Moral Sentiments in Hume's *Treatise*. *The Review of Metaphysics* 30(1): 57-72.

Hobbes, Thomas. (1994). *Leviathan.* Ed. E. Curley. Indianapolis: Hackett ('L').

Horgan, Terry and Mark Timmons. (2006). Cognitivist Expressivism. In *Metaethics after Moore*, ed. Terry Horgan and Mark Timmons, 255-298. Oxford: Clarendon Press.

———. (2007). Morphological Rationalism and the Psychology of Moral Judgment. *Ethical Theory and Moral Practice* 10(3): 279-295.

Hume, David. (1978). *A Treatise of Human Nature*. Ed. L.A. Selby-Bigge, revised by P.H. Nidditch (2nd edition). Oxford: Oxford University Press ('T').

———. (1975). *Enquiries concerning Human Understanding and concerning the Principles of Morals*. Ed. L.A. Selby-Bigge, revised by P.H. Nidditch (3rd edition). Oxford: Oxford University Press ('E', 'M').

———. (2007). *A Dissertation on the Passions and the Natural History of Religion*. Ed. Tom L. Beauchamp. Oxford: Clarendon Press ('P').

———. (1987). *Essays, Moral, Political, and Literary*. Ed. Eugene F. Miller, Eugene. Indianapolis: Liberty Fund ('EMPL').

Hutcheson, Francis. (1971). *Illustrations on the Moral Sense*. Cambridge MA: The Belknap Press of Harvard University Press.

Jackson, Frank and Philip Pettit. (1998). A Problem for Expressivism. *Analysis* 58(4): 239-251.

Jackson, Frank, Philip Pettit and Michael Smith. (2000). Ethical Particularism and Patterns. In *Moral Particularism*, ed. Brad Hooker and M.O. Little, 79-99. Oxford: Clarendon Press.

Kalderon, Mark. (2005). *Moral Fictionalism*. Oxford: Clarendon Press.

Kemp Smith, Norman. (1966). *The Philosophy of David Hume*. London: MacMillan and Company Ltd.

Kennett, Jeanette and Cordelia Fine. (2009). Will the Real Moral Judgment Please Stand Up? *Ethical Theory and Moral Practice* 12(1): 77-96.

Köhler, Sebastian. (2012). Expressivism, Subjectivism and Moral Disagreement. *Thought* 1: 71-78.

Korsgaard, Christine. (1999). The General Point of View: Love and Moral Approval in Hume's Ethics. *Hume Studies* 25(1/2): 3-42.

Kriegel, Uriah. (2012). Moral Motivation, Moral Phenomenology, and the Alief/Belief Distinction. *Australasian Journal of Philosophy* 90(3): 469-486.

Krause, Sharon R. (2004). Hume and the (False) Luster of Justice. *Political Theory* 32(5): 628-655.

Landy, David. (2006). Hume's Impression/Idea Distinction. *Hume Studies* 32(1): 119-139.

———. (2012). Hume's Theory of Mental Representation. *Hume Studies* 38(1): 23-54.

Lebrecht, Sophie, Moshe Bar, Lisa Feldman-Barrett and Michael J. Tarr. (2012). Micro-valences: perceiving affective valence in 'neutral' everyday objects. *Frontiers in Perception Science* 3: 1-5.

Lenman, James. (1999). Michael Smith and the Daleks: Reason, Morality, and Contingency. *Utilitas* 11(2): 164-177.

Lewis, David. (1989), Dispositional Theories of Value II. *Proceedings of the Aristotelian Society, Supplementary Volumes* 63(1): 113-137.

Locke, John. (2008). *An Essay Concerning Human Understanding*. Ed. Pauline Phemister. Oxford: Oxford University Press.

Loeb, Louis E. (1977). Hume's Moral Sentiments and the Structure of the *Treatise*. *Journal of the History of Philosophy* 15(4): 395–403.

———. (2005). *Stability and Justification in Hume's Treatise*. New York: Oxford University Press.

Lowe, E.J. (2006). Ideational Theories of Meaning. In *Concise Encyclopedia of Philosophy of Language and Linguistics*, eds. Alex Barber and Robert J. Stainton, 299-301. Oxford: Elsevier.

Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. London: Penguin Books.

———. (1980). *Hume's Moral Theory*. Oxon: Routledge.

Magri, Tito. (2008). Hume on the Direct Passions and Motivation. In *A Companion to Hume*, ed, , Elizabeth S. Radcliffe, 185-200. Malden, MA: Blackwell Publishing Ltd.

McDowell, John. (1981). Following a Rule and Ethics. In *Wittgenstein: To Follow a Rule*, eds. S. Holtzman and Christopher M. Leich, 141-162. London: Routledge.

McIntyre, Jane L. (2000). Hume's Passions: Direct and Indirect. *Hume Studies* 26(1): 77-86.

Mercer, Philip. (1972). *Sympathy and Ethics*. Oxford: Clarendon Press.

Merivale, Amyas. (2019). *Hume on Art, Emotion, and Superstition: A Critical Study of the Four Dissertations.* New York: Routledge.

Miller, Alexander. (2005). *An Introduction to Contemporary Metaethics*. Cambridge: Polity Press.

Millgram, Elijah. (1995). Was Hume a Humean? *Hume Studies* 21(1): 75-93.

Millican, Peter. (2017). Hume's Fork and his Theory of Relations. *Philosophy and Phenomenological Research* 95(1): 3-65.

Moore, G.E. (1903). *Principia Ethica*. Cambridge: Cambridge University Press.

———. (2005). *Ethics: The Nature of Moral Philosophy*. Oxford: Clarendon Press.

Moore, James. (1994). Hume and Hutcheson. In *Hume and Hume's Connexions*, edited by M.A. Stewart and John P. Wright, 23-57. Edinburgh: Edinburgh University Press.

Nagel, Jennifer. (2014). Intuition, Reflection and the Command of Knowledge. *Proceedings of the Aristotelian Society Supplementary Volume* 88(1): 217-240.

Noonan, Harold. (1999). *Hume on Knowledge.* London: Routledge.

Norton, David Fate. (1982). *David Hume: Common-Sense Moralist, Sceptical Metaphysician*. Princeton: Princeton University Press.

Owen, David. (2002). *Hume's Reason*. Oxford: Oxford University Press.

Paxman, Katharina. (2015). Imperceptible Impressions and Disorder in the Soul: A Characterization of the Distinction between Calm and Violent Passions in Hume. *The Journal of Scottish Philosophy* 13(3): 265-278.

Penelhum, Terence. (1975). *Hume*. London and Basingstoke: The MacMillan Press.

Persson, Ingmar. (1997). Hume – Not a 'Humean' about Motivation. *History of Philosophy Quarterly* 14(2): 189-206.

Pigden, Charles. (2007). Hume, Motivation and 'The Moral Problem'. In *New Essays on David Hume*, eds. Emilio Mazza and Emanuele Ronchetti, 199-221. Milano, Franco Angeli.

Pölzler, Thomas and Jennifer Cole Wright. (2019). Empirical research on folk moral objectivism. *Philosophy Compass* 14(5): 1-15.

———. (2020). Anti-Realist Pluralism: A New Approach to Folk Metaethics. *Review of Philosophy and Psychology* 11(1): 53–82.

Prinz, Jesse. (2007). *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Qu, Hsueh. (2012). The Simple Duality: Humean Passions. *Canadian Journal of Philosophy* 42(S1): 98–116.

Radcliffe, Elizabeth S. (1994) Hume on Motivating Sentiments, the General Point of View, and the Inculcation of 'Morality'. *Hume Studies* 20(1): 37-58.

———. (1999). Hume on the Generation of Motives: Why Beliefs Alone Never Motivate. *Hume Studies* 25(1/2): 101-122.

———. (2004). Love and Benevolence in Hutcheson's and Hume's Theories of the Passions. *British Journal for the History of Philosophy* 12(4): 631-653.

———. (2015a). Hume's Psychology of the Passions: The Literature and Future Directions. *Journal of the History of Philosophy* 53(4): 565-605.

———. (2015b). Strength of Mind and the Calm and Violent Passions. *Res Philosophica* 92(3): 547–567.

Ridge, Michael. (2014). *Impassioned Belief*. Oxford: Oxford University Press.

Reed, Philip A. (2016) Hume on Sympathy and Agreeable Qualities. *British Journal for the History of Philosophy* 24(6): 1136-1156.

Roberts, Debbie. (2011). Shapelessness and the Thick. *Ethics* 121(3): 489-520.

Ross, W.D. (1930). *The Right and the Good*. Oxford: Clarendon Press.

Sarkissian, Hagop, John Park, David Tien, Jennifer Cole Wright and Joshua Knobe. (2011). Folk Moral Relativism. *Mind & Language* 26(4): 482-505.

Sayre-McCord, Geoffrey. (1994). On Why Hume's General Point of View Isn't Ideal – and Shouldn't Be. *Social Philosophy and Policy* 11(1): 202-228

———. (2008). Hume on Practical Morality and Inert Reason. In *Oxford Studies in Metaethics: Vol. 3*, ed. Russ Shafer-Landau, 299-320. Oxford: Oxford University Press.

Schauber, Nancy. (1999). Hume on Moral Motivation: It's Almost like Being in Love. *History of Philosophy Quarterly*, 16(3): 341- 366.

Schein, Chelsea and Kurt Gray. (2015). The unifying moral dyad: liberals and conservatives share the same harm-based template. *Personality and Social Psychology Bulletin*. 41(8): 1147–1163.

———. (2018). The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*. 22(1): 32–70.

Schmitter, Amy M. (2008). Making an Object of Yourself: On the Intentionality of the Passions in Hume. In *Topics in Early Modern Philosophy of Mind*, ed. J. Miller, 223-240. Dordrecht: Springer.

Schneewind, J.B. (1998). *The Invention of Autonomy*. Cambridge: Cambridge University Press.

Schroeder, Mark. (2007). *Slaves of the Passions*. Oxford: Oxford University Press.

———. (2008). What is the Frege-Geach Problem? *Philosophy Compass* 3(4): 703–720.

———. (2009). Hybrid Expressivism: Virtues and Vices. *Ethics* 119(2): 257–309.

Shafer-Landau, Russ. (2003). *Moral Realism: A Defence*. Oxford: Oxford University Press.

Sinclair, Neil. (2008). Free Thinking for Expressivists. *Philosophical Papers* 37(2): 263-287.

———. (2012). Moral Realism, Face-Values and Presumptions. *Analytic Philosophy* 53(2): 158-179.

———. (2014). On Standing One's Ground. *Analysis* 74(3): 422-431.

———. (2016). Reasons, Inescapability and Persuasion. *Philosophical Studies* 173(10): 2823-2844.

———. (2020). *Ethical Subjectivism and Expressivism*. Cambridge: Cambridge University Press.

Sinhababu, Neil. (2017). *Humean Nature: How desire explains action, thought, and feeling*. Oxford: Oxford University Press.

Sinnott-Armstrong, Walter. (2008). Is moral phenomenology unified? *Phenomenology and the Cognitive Sciences*. 7(1): 85–97.

Sinnott-Armstrong, Walter and Thalia Wheatley. (2014). Are moral judgments unified? *Philosophical Psychology*. 27(4): 451-474.

Sinnott-Armstrong, Walter, Liane Young and Fiery Cushman. (2010). Moral Intuitions. In *The Moral Psychology Handbook*, ed. John M. Doris and the Moral Psychology Research Group, 246-272. Oxford: Oxford University Press.

Sloman, Steven. (1996) The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin* 119(1): 3–22.

Slote, Michael. (2010). *Moral Sentimentalism*. Oxford: Oxford University Press.

Smith, Michael. (1994). *The Moral Problem*. Oxford: Blackwell Publishers Ltd.

Snare, Frank. (1975). The Argument from Motivation. *Mind* 84(333): 1-9.

Stevenson, Charles Lewis. (1944), *Ethics and Language*. New Haven: Yale University Press.

———. (1963). *Facts and Values: Studies in Ethical Analysis*. New Haven: Yale University Press.

Strandberg, Caj. (2012). A Dual Aspect Account of Moral Language. *Philosophy and Phenomenological Research* 84(1): 87-122.

Stroud, Barry. (1977). *Hume*. Oxon: Routledge and Kegan Paul plc.

Suikkanen, Jussi. (2009). The Subjectivist Consequences of Expressivism. *Pacific Philosophical Quarterly* 90(3): 364-387.

Sweigart, John. (1964). The Distance Between Hume and Emotivism. *Philosophical Quarterly* 14(56): 229-236.

Taylor, Jacqueline. (1998). Justice and the Foundations of Social Morality in Hume's *Treatise*. *Hume Studies* 24(1): 5-30.

———. (2002). Hume on the Standard of Virtue. *The Journal of Ethics* 6(1): 43-62.

———. (2013). Hume on the Importance of Humanity. *Revue Internationale de Philosophie* 263(1): 81-97.

———. (2015). *Reflecting Subjects.* Oxford: Oxford University Press.

Vitz, Rico. (2004). Sympathy and Benevolence in Hume's Moral Psychology. *Journal of the History of Philosophy* 42(3): 261-275.

Waxman, Wayne. (2003). *Hume's Theory of Consciousness.* Cambridge: Cambridge University Press.

Williams, Bernard. (1985). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Wright, Jennifer C., Piper T. Grandjean and Cullen B. McWhite. (2013). The meta-ethical grounding of our moral beliefs: Evidence for meta-ethical pluralism. *Philosophical Psychology*, 6(3): 336–361.

Zajonc, Robert. B., and Hazel Markus. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research* 9(2): 123–131.

Zangwill, Nick. (1990). Quasi-Quasi-Realism. *Philosophy and Phenomenological Research*. 50(3): 583-594.